Validity evaluation of the Lawson classroom test of scientific reasoning

Lei Bao,[1,*] Yang Xiao,[1,2] Kathleen Koenig,[3] and Jing Han[1]
[1]*The Ohio State University, Columbus, Ohio 43210, USA*
[2]*South China Normal University, Guangzhou, Guangdong 510631, China*
[3]*University of Cincinnati, Cincinnati, Ohio 45221, USA*

(Received 15 June 2018; published 17 August 2018)

In science, technology, engineering, and mathematics education there has been increased emphasis on teaching goals that include not only the learning of content knowledge but also the development of scientific reasoning skills. The Lawson classroom test of scientific reasoning (LCTSR) is a popular assessment instrument for scientific reasoning. Through large scale applications, however, several issues have been observed regarding the validity of the LCTSR. This paper will review the literature on the assessment of scientific reasoning and provide a detailed analysis of the current version of LCTSR in regards to its validity and measurement features. The results suggest that the LCTSR is a practical tool for assessing a unidimensional scale of scientific reasoning. The instrument has a good overall reliability for the whole test. However, inspections of individual question pairs reveal a variety of validity concerns for five of the total twelve question pairs. These five question pairs have a considerable amount of inconsistent response patterns. Qualitative analysis also indicates wide ranging question design issues. Therefore, assessment of subskills involved with the five question pairs may have less power due to their questionable validity.

DOI: 10.1103/PhysRevPhysEducRes.14.020106

I. INTRODUCTION

In science, technology, engineering, and mathematics (STEM) education, widely accepted teaching goals include not only the learning of broad content knowledge, but also the development of general scientific reasoning abilities. These abilities are considered necessary for contributing to the workforce and global economy of the 21st century [1]. Scientific reasoning, which is closely related to "formal operational reasoning" [2] and "critical thinking" [3], represents the thinking and reasoning skills involved during inquiry, experimentation, evaluation of evidence, inference, and argument that support the formation and modification of concepts and theories about the natural and social world [4–6]. Early research suggests that scientific reasoning abilities can be developed through training and can be transferred to support learning in other areas [7,8]. Training in scientific reasoning abilities can also have a long-term impact on students' academic achievement, making their value broadly appealing rather than subject limited [7].

More recently, there has been increased interest in researching student abilities in scientific reasoning. For example, Bao *et al.* [4] reported that the traditional style of STEM education has little impact on the development of students' scientific reasoning abilities. Meanwhile, correlational studies have found that scientific reasoning abilities positively correlate with course achievement [9,10], performances on concept tests [11,12], and success on transfer reasoning questions [13,14]. Such results prompt the need for more widespread and proactive approaches in improving educational environments to more actively target scientific reasoning. To move forward, educators and researchers need good assessment tools that can be easily applied in large scale settings and produce valid results for evaluating scientific reasoning abilities. This is important for gauging the abilities of students such that appropriate skill level teaching methods can be implemented, as well as for evaluating students' learning gains in scientific reasoning under different educational settings.

Historically, the Piagetian clinical interview was one of the early methods used to assess students' formal reasoning abilities. Such a method is cost and time intensive, however, making it difficult for classroom practices [15,16]. Guided by Piagetian tasks, a number of researchers developed their own instruments in assessing student scientific reasoning abilities, such as the group assessment of logical thinking test (GALT) [17], the test of logical thinking (TOLT) [18], and Lawson's classroom test of

*Corresponding author.
bao.15@osu.edu

formal reasoning (CTFR-78) [16]. The initial open-response version of the CTFR-78 was revised in the year 2000 to become a 24-item multiple choice test [19]. This most recent version is referred to as the Lawson classroom test of scientific reasoning (LCTSR), and it has gained wide popularity in the STEM education community. For example, physics education researchers used the LCTSR to study the relation between student scientific reasoning abilities and physics content learning. In one study, Coletta and Phillips [11] reported correlations ($r \approx 0.5$) between pre-post normalized gains on the Force Concept Inventory [20] and student reasoning abilities.

Although several studies [21,22] investigated the validity of the CTFR-78, research on the validity of the LCTSR is limited. Nevertheless, the instrument has become a standard assessment tool in education research. Through our own use of the LCTSR, however, several issues have been observed concerning question design and data interpretation. This study evaluates the validity of the LCTSR based on large-scale assessment data and follow-up interviews. The findings will further establish the validity of the LCTSR and reveal measurement features of the current design.

II. LITERATURE REVIEW ON ASSESSMENT OF SCIENTIFIC REASONING

A. The development and validation of the CTFR-78

During the 1960s and 1970s, research on the assessment of formal reasoning led to the development of several instruments, which employed paper-and-pencil methods that could be more conveniently administered and scored than clinical interviews [23–26]. A few of the instruments required students to interact with equipment during the assessment, but responses to the questions were in the form of written response [27]. However, the use of this type of instrument was limited due to the equipment needs, and studies that depended on this instrument tended to have smaller sample sizes. Building on this idea, other instruments were developed that simplified the process, and an instructor was used to conduct a demonstration for the entire class to minimize time and equipment requirements [28]. This method sought balance between the power of interviews and a format that could be more readily implemented.

One of the more popular assessments during this period was the 1978 version of Lawson's classroom test of formal reasoning (CTFR-78) [16]. The CTFR-78 was developed based on Piaget's developmental theory [2] and utilized questions that others had created in different contexts to test formal reasoning for a variety of operations [17,18,26]. The CTFR-78 involves fifteen items that each consist of a demonstration conducted by an instructor, followed by two questions that students answer in individual test booklets. The first question is multiple choice and asks for outcomes related to the demonstration. The second question is a written free-response question, in which students explain the reasoning for their answer to the first question. An item is scored as correct only if answers to both questions are deemed satisfactory. The test covers five skill dimensions including conservation, proportion, control of variables, combination, and probability (see Table I).

In a paper that introduced the CTFR-78, Lawson [16] conducted a series of studies to establish the initial validity of the CTFR-78 in measuring concrete and formal reasoning. In the study, six experts in Piagetian research assessed the quality of the questions; all six agreed that the questions assessed concrete and formal reasoning. Subsequently, the instrument was administered to 513 students from eighth through tenth grade, and 72 of these students were randomly selected to participate in a series of clinical interviews with Piagetian tasks. Two statistical methods, parametric statistics and principal components analysis, were employed for comparing the interview and test data. The former method found an overall correlation of 0.76 ($p < 0.001$). The principal components analysis identified three major principal factors among the five skill dimensions that accounted for 66% of the variance in scores. Because of the small number of students, this result was considered tentative but satisfactory. Overall, the analysis indicated that the CTFR-78 was able to measure formal reasoning and that it correlated reasonably well with clinical methods [16].

Lawson also created a scoring system to help instructors interpret test results regarding students' levels on a Piagetian reasoning scale. Based on the test scores of 513 students, Lawson defined three levels of reasoning: concrete (score of 0–5), transitional (score of 6–11), and formal (score of 12–15) [16]. Roughly one-third (35.3%) of students examined were classified at the concrete level, about half (49.5%) transitional, and the remaining (15.2%) at the formal level. When compared to interview assessments, more than half of the 72 interviewees fit into the same three reasoning levels determined by the CTFR-78. Of those that did not fit, the data indicated that the CTFR-78 may have underestimated the reasoning abilities of the students.

Other researchers also analyzed the CTFR-78 to determine its ability to measure student reasoning. Stefanich *et al.* [22] measured the correlation between outcomes of clinical interviews with the CTFR-78, and observed a weaker correlation ($r = 0.50$) than that reported by Lawson. Interestingly, when compared to the clinical interviews, the researchers found that the test overestimated reasoning abilities rather than underestimated, but due to a small sample size ($N = 27$) with no reported estimate of statistical significance, the study did little to challenge the earlier findings of Lawson.

Another study by Pratt and Hacker [21] involved administering the test to 150 students to uncover whether the CTFR-78 measured one or several factors. The researchers found that the Lawson test measurement was multifactorial rather than singular, which they took to be a

TABLE I. The comparison of CTFR-78 and LCTSR.

Scheme tested	Items (CTFR-78)	Question pair (LCTSR)	Task details
Conservation of weight	1	1, 2	Varying the shapes of two identical balls of clay placed on opposite ends of a balance.
Conservation of volume	2	3, 4	Examining the displacement volumes of two cylinders of different densities.
Proportional reasoning	3, 4	5, 6, 7, 8	Pouring water between wide and narrow cylinders and predicting levels.
Proportional reasoning	5, 6		Moving weights on a beam balance and predicting equilibrium positions.
Control of variables	7	9, 10	Designing experiments to test the influence of length of string on the period of a pendulum.
Control of variables	8		Designing experiments to test the influence of weight of bob on the period of a pendulum.
Control of variables	9, 10		Using a ramp and three metal spheres to examine the influence of weight and release position on collisions.
Control of variables		11, 12, 13, 14	Using fruit flies in tubes to examine the influence of red/blue light and gravity on flies' responses.
Combinational reasoning	11		Computing combinations of four switches that will turn on a light.
Combinational reasoning	12		Listing all possible linear arrangements of four objects representing stores in a shopping center.
Probability	13, 14, 15	15, 16, 17, 18	Predicting chances for withdrawing certain colored wooden blocks from a sack.
Correlation reasoning		19, 20	Predicting whether correlation exits between the size of the mice and the color of their tails through presented data.
Hypothetical-deductive reasoning		21, 22	Designing experiments to determine why the water rushed up into the glass after the lit candle went out.
Hypothetical-deductive reasoning		23, 24	Designing experiments to determine why red blood cells become smaller after adding a few drops of salt water.

weakness of the test. This examination was later repeated by Hacker [29] who again found the test to be multifactorial. Other researchers [30] did not find a multifactorial examination to be problematic, especially given that formal reasoning is multifaceted.

And last, another early study investigated how strongly the CTFR-78 test scores correlated to student success in science and mathematics [31]. It was found that students at the "formal operational reasoning" level outperformed students at the "transitional" and "concrete operational reasoning" levels in sciences and mathematics, though the latter two levels of ability were indistinguishable. Concerning general performance in STEM, only one of the items (probabilistic reasoning) was found to be predictive, but overall the test was a good indicator of success in learning biology but is less effective in predicting success in other STEM areas.

B. The development and validation of the LCTSR

Two decades later, the LCTSR was published, which is a completely multiple-choice assessment and does not involve demonstrations. This newer version uses a two-tier design that has a total of 24 questions in 12 pairs and contains a total of six skill dimensions. Items on combinational reasoning in the CTFR-78 were not included in the new LCTSR, while new items were added for correlational reasoning and hypothetical-deductive reasoning. A comparison of CTFR-78 and LCTSR is provided in Table I. In the discussion that follows, the terms "item" and "question" will be used interchangeably based on the styles of the relevant literature.

The first 10 question pairs of the LCTSR maintained the traditional two-tier format, in which the first question in a pair asks for the results of a given task, and the second question asks for the explanation and reasoning behind the answer to the first question. The last two pairs (21-22 and 23-24) on hypothetical-deductive reasoning have a slightly changed flavor of paired questions. Question 21 asks students to determine the most appropriate experimental design that can be used to test a given hypothesis for a provided phenomenon. Question 22 asks students to pick the experimental outcome that would disprove the hypothesis proposed in question 21. Questions 23-24 are similarly structured and ask students to select the experimental outcomes that will disprove two given hypotheses.

The scoring of the LCTSR has evolved into two forms: one is the pair scoring method, which assigns one point when both questions in a pair are answered correctly, and the other is the individual scoring method, which simply grades each question independently. Both methods have been frequently used by researchers [4,11,16,32–34].

The utility of the LCTSR is that it can be quickly and objectively scored, and thus it has become widely used. However, the validity of this new version has not been thoroughly evaluated, but rather rests on its predecessor. Regardless, many researchers have used the LCTSR and compared the results with other assessments. For example, small but statistically significant correlations have been demonstrated between LCTSR test scores and changes of pre-post scores on both the Force and Concept Inventory (FCI) ($r = 0.36$, $p < 0.001$) and Conceptual Survey of Electricity and Magnetism (CSEM) ($r = 0.37$, $p < 0.001$) among community college students, including those who have and have not taken a calculus course [35]. Similar findings have been reported by others [11,32,36], indicating that the LCTSR is a useful measure of formal reasoning.

Although widely used, a complete and formal validation, including construct, content, and criterion validity of the LCTSR, has not yet been conducted [37]. Only recently have a few studies started to examine its construct validity [38–41]. For example, using Rasch analysis and principle component analysis, the assumption of a unidimensional construct for a general scientific reasoning ability has been shown to be acceptable [38,39]. In addition, the multidimensionality of LCTSR has been partially established using confirmatory factor analysis [40]. As a two-tier instrument, the LCTSR is assumed to measure students' *knowing* in the first tier and their *reasoning* processes in the second tier [42], and this has been confirmed in a recent empirical study [41].

C. The two-tier multiple-choice design

A two-tier multiple-choice (TTMC) design is constructed to measure students' content knowledge in tier 1 and reasoning in tier 2 [42,43]. Since its formal introduction three decades ago [42], TTMC designs have been widely researched and applied in education assessment [44–52]. The LCTSR is a well-known example of a two-tier test designed to assess scientific reasoning [11,16,41,42,44].

Two assessment goals are often assumed when using TTMC tests. The first is to measure students' knowing and reasoning at the same time within a single question setting. Understanding the relation between knowing and reasoning can provide important cognitive and education insights into how teaching and learning can be improved. Recent studies have started to show a possible progression from knowing to reasoning, which suggests that reasoning is harder than knowing [34,41,53–56]. However, the actual difficulty reflected within a given test can also be affected by a number of factors including test design, content, and population [34,41], and therefore needs to be evaluated with all the possible factors considered and or controlled.

The second, and often more primary goal of using the TTMC design, is to suppress "false positives" when the pair scoring method is used. It is possible for students to answer a question correctly without the correct understanding through guessing or other means. From a signal processing perspective, this type of "positive" signal is generated with unfavorable processes, and was defined as a false positive by Hestenes and Halloun [57]. For example, in the case of the Force Concept Inventory (FCI) [20], which has five answer choices for each question, the chance of a false positive due to guessing can be estimated as 20% [58]. Using TTMC or other sequenced question design, a false positive can be suppressed in pair or sequence scoring [51]. For example, the false positive due to guessing drops to 4% in pair scoring for the five-choice question pair ($1/5 \times 1/5 = 1/25$).

In contrast, it is also possible for a student to answer a question incorrectly but with correct understanding or reasoning, which was defined as a "false negative" by Hestenes and Halloun [57]. It has been suggested that a false negative is less common than a false positive in a well-designed instrument, such as the FCI [20]. However, when the test designs are subject to inspection, both false positives and false negatves should be carefully considered and examined. For a well-designed TTMC instrument, good consistency between answer and explanation is generally expected. If a high level of inconsistency is observed, such as a correct answer for a wrong reason or vice versa, it could be an indication of content validity issues in question design.

D. Research question

The construct validity of the LCTSR has been partially established through quantitative analysis, but there has been little research on fully establishing its content validity. Based on large scale implementations of the LCTSR, a number of concerns regarding question designs have been raised, prompting the need for a thorough inspection of the assessment features of the LCTSR and to fully evaluate its validity. This will allow researchers and educators to more accurately interpret the results of this assessment instrument. Therefore, this paper aims to address a gap in the literature through a detailed study of the possible content validity issues of LCTSR, such that the results will help formally establish the validity of this popular instrument. This study focuses on one core research question: to what extent is LCTSR valid and reliable to measure scientific reasoning?

III. METHODOLOGY

A. Data collection

This research uses mixed methods [59] to integrate different sources of data to identify and study features of

test design and validity of the LCTSR. Three forms of quantitative and qualitative data were collected from freshman in three midwestern public universities including the following: (i) large-scale quantitative data ($N = 1576$) using the LCTSR as written; (ii) student responses ($N = 181$) to LCTSR questions along with short answer free-response explanations for each question; and (iii) think-aloud interviews ($N = 66$) with a subgroup randomly selected from the same students tested in part (ii). During the interviews, students were asked to go over the test again and verbally explain their reasoning for how they answered the questions.

Among the students in the large-scale quantitative data, 988 were enrolled in calculus-based introductory physics courses, with the rest enrolled in algebra-based introductory physics courses. The universities selected were those with medium ranking in order to obtain a representative pool of the population. For the college students, the LCTSR was administered in the beginning of the fall semester prior to any college level instruction.

In addition to the student data, an expert review panel consisting of 7 science faculty (3 in physics, 2 in biology, 1 in chemistry, and 1 in mathematics) was formed to provide expert evaluation of the questions for content validity. These faculty are researchers in science education with established experience in assessment design and validation.

B. Analysis method

1. The dependence within and between question pairs

In general, test design features can result in dependences within and between question pairs. For example, in large-scale testing programs such as TOEFL, PISA, or NAEP, there are frequently observed correlations among questions sharing common context elements such as graphs or texts. These related question sets are often called testlets [60] or question bundles [61]. In this way, a typical two-tier question pair can be considered as a specific type of testlet or question bundle [62]. In addition, the LCTSR was designed with six subskill dimensions that each can contain multiple question pairs built around similar contextual scenarios. As a result, dependences between question pairs within a single skill dimension are not surprising. The analysis of correlations among related questions can provide supporting evidence for identifying and validating factors in question design that might influence students' performances one way or the other.

In classical score-based analysis, dependences among questions are analyzed using correlations among question scores. This type of correlation contains contributions from all factors including students' abilities, skill dimensions, question difficulties, question design features, etc. If the data are analyzed with Rasch or other item response theory (IRT) models, one can remove the factors due to students' abilities and item difficulties and focus on question design features and other factors. Therefore, the fit measure Q_3 [63], which gives the correlation among question residuals [Eqs. (1) and (2)], is used to analyze the dependences among questions in order to examine the influences on students' performance variances from question design features:

$$d_{ik} = X_{ik} - \hat{P}_i(\hat{\theta}_k). \tag{1}$$

Here d_{ik} is the difference (residual) between X_{ik}, the observed score of the kth student on the ith question, and the model predicted score $\hat{P}_i(\hat{\theta}_k)$. The Pearson correlation of these question residuals is computed as

$$Q_{3,ij} = r_{d_i d_j}, \tag{2}$$

where d_i and d_j are question residuals for questions i and j, respectively. In using Q_3 to screen questions for local dependence, Chen and Thissen [64] suggested an absolute threshold value of 0.2, above which a moderate to strong dependence can be derived. For a test consisting of n questions, the mean value of all Q_3 correlations among different questions is expected to be $-1/(n-1)$ [65]. The LCTSR contains 24 questions; therefore, the expected mean Q_3 is approximately -0.043.

The LCTSR has a TTMC design with six subskills. By design, there are at least two categories of dimensions including the two-tier structure and content subskills. Additional observed dimensions beyond the two categories can be considered as a result of contextual factors in question design. In data analysis, a multidimensional Rasch model is used to account for the dependences due to the subskill dimensions. Then, the variances of the residuals [Eq. (1)] would primarily be the results of TTMC structure and other question design features. For an ideal TTMC test, it is often expected that dependences only exist within, but not between, question pairs. This is equivalent to the condition of local independence among question pairs. With the subskill dimensions removed by the multidimensional Rasch modeling, significant residual dependences among different question pairs can be indications of possible construct and content validity issues in the question design.

2. Pattern analysis of the LCTSR

As discussed earlier, the first 10 LCTSR question pairs (questions 1 through 20) used the traditional two-tier design, while the last two pairs (questions 21-22 and 23-24) have a different flavor of paired questions. Ideally, traditional two-tier questions would expect good consistency between tier-1 and tier-2 questions (i.e., between answer and explanation). If a high level of inconsistency is observed, such as a correct answer for a wrong reason or vice versa, it may suggest issues in question design that can lead to false positive or negative assessment outcomes. Therefore, analysis of question-pair

TABLE II. Question-pair score response patterns. A correct answer is assigned with 1 and a wrong answer is assigned with 0.

Response patterns	Description	Scoring	
		Pair	Individual
11	Consistent correct, correct answer with correct reasoning	1	2
00	Consistent wrong, wrong answer with wrong reasoning	0	0
10	Inconsistent, correct answer with wrong reasoning	0	1
01	Inconsistent, wrong answer with correct reasoning	0	1

response patterns can provide useful information to identify potential content validity issues. For a TTMC question pair, there are four score-based question pair response patterns (11, 00, 10, 01), which are listed in Table II. Patterns 00 and 11 represent consistent answers, while patterns 10 and 01 are inconsistent answers.

In this study, the inconsistent patterns were analyzed in detail to help identify possible content validity issues of certain questions. For each question pair, the frequencies of the different response patterns from the entire population were calculated and compared with other question pairs regarding the popularity of inconsistent patterns. Question pairs with a significantly higher level of inconsistent patterns were identified for further inspection on their question designs for validity evaluation.

In addition, for a specific question pair, students' responses should vary with their performance levels. Students with high test scores are often expected to be less likely to respond with inconsistent patterns [66,67]. When the measurement results on certain question pairs depart from such expectations, indications of content validity issues can be expected. To study this type of trending relations, distributions of inconsistent patterns from low to high score levels were calculated and compared. For the ith question pair, probability $P_i(s)$ was defined to represent the normalized frequency of the inconsistent patterns for the subgroup of students with score s. Using individual scoring, students' test scores on the LCTSR were binned into 25 levels of measured scores ($s = 0, 1, 2, \ldots, 24$). For each binned score level, one can calculate $P_i(s)$ with

$$P_i(s) = \frac{N_i(s)}{N(s)}. \tag{3}$$

Here $N(s)$ is the total number of students with score s, while $N_i(s)$ is the number of students out of $N(s)$ who have inconsistent responses on the ith question pair. The overall P_i and its distribution of $P_i(s)$ over s were analyzed for each question pair. Question pairs with large P_i and/or abnormal distributions (e.g., large P_i at high s) were identified for further inspection on possible validity issues.

Furthermore, using the individual scoring method the scores gained from inconsistent responses were compared to the total scores to evaluate the impact of the inconsistent responses on score-based assessment. Two measures were used: one is denoted as $R_{12}(s)$ to represent the fraction of scores from inconsistent responses out of all 12 question pairs for students with score s, and the other is $R_5(s)$, which gives the fraction of scores from inconsistent responses out of 5 question pairs identified as having significant validity issues (to be discussed in later sections). The two fractions were calculated using Eqs. (4) and (5):

$$R_{12}(s) = \frac{\sum_{i=1}^{12} P_i(\mathrm{s})}{\mathrm{s}}, \tag{4}$$

$$R_5(s) = \frac{\sum_5 P_i(\mathrm{s})}{\mathrm{s}}. \tag{5}$$

3. Qualitative analysis

The analysis was conducted on three types of qualitative data including expert evaluations, student surveys with open-ended explanations, and student interviews. The results were synthesized with the quantitative analysis to identify concrete evidence for question design issues of LCTSR and their effects on assessment results.

IV. RESULTS OF QUANTITATIVE ANALYSIS TO IDENTIFY POSSIBLE VALIDITY ISSUES

A. Basic statistics

The distributions of students' scores using both pair and individual scoring methods were calculated and plotted in Fig. 1. For this college population, the distributions appear to be quite similar and both are skewed to the higher end of the scale. On average, the mean test score from the pair scoring method (mean = 58.47, SD = 23.03%) is significantly lower than the score from the individual scoring method (mean = 68.03, SD = 20.33%) ($p < 0.001$, Cohen's $d = 1.682$). The difference indicates that there exists a significant number of inconsistent responses (see Table II), which will be examined in the next section.

The reliability of LCTSR was also computed using scores from both scoring methods. The results show that the reliability of the pair scoring method ($\alpha = 0.76$, n of question paris = 12) is slightly lower than that of the individual scoring method ($\alpha = 0.85$, n of questions = 24) due to the smaller number of scoring units with pair scoring. To counter for the difference in test length, reliability of the

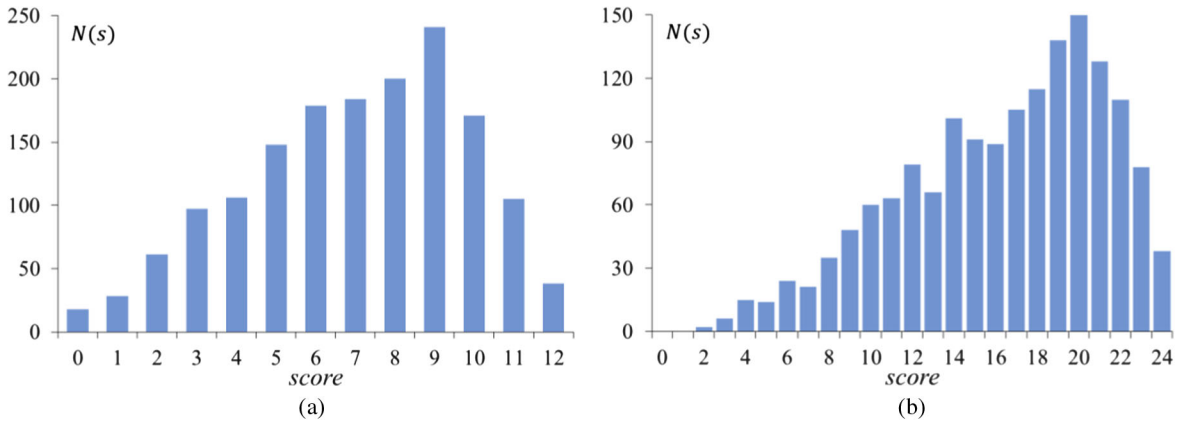

FIG. 1. Score distribution: (a) pair scoring method and (b) individual scoring method.

pair scoring method was adjusted by the Spearman-Brown prophecy formula, which gives a 24-item equivalent α at 0.86, almost identical to that of the individual scoring method. The results suggest that the LCTSR produces reliable assessment scores with both scoring methods [68]. The reliabilities of the 6 subskills were also computed using the individual scoring method. The results show that the reliabilities of the first 5 subskills are acceptable ($\alpha = 0.71$, 0.81, 0.65, 0.82, and 0.83, respectively) [68]. For the last subskill (HD), the reliability is lower than the acceptable level ($\alpha = 0.52 < 0.65$).

Using the individual scoring method, the item mean, discrimination, and point biserial coefficient for each question were computed and listed in Table III. The results show that most item means, except for question 1, 2, and 16, fall into the suggested range of difficulty [0.3, 0.9] [69]. These three questions appear to be too easy for the college students tested, which also result in poor discrimination (<0.3) [69]. For all questions, the point biserial correlation coefficients are larger than the criterion value of 0.2, which suggests that all questions hold consistent measures with the overall LCTSR test [69].

The mean score for each question was also plotted in Fig. 2. The error bars represent the standard error of the mean, which are very small due to the large sample size. When the responses to a question pair contain a significant number of inconsistent patterns, the scores of the two individual questions are less correlated. The mean scores of the two questions in the pair may also be substantially different but are dependent on the relative numbers of the 10 and 01 patterns. To compare among the question pairs, the correlation and *effect size* (Cohen's d) between question scores within each question pair were calculated and plotted in Fig. 2. The results show that among the 12 question pairs, seven (1-2, 3-4, 5-6, 9-10, 15-16, 17-18, and 19-20) have high correlations ($r > 0.5$) and relatively small differences in score. The remaining five pairs (7-8, 11-12, 13-14, 21-22, and 23-24) have considerably smaller correlations ($r < 0.4$) and larger differences in score, indicating a more prominent presence of inconsistent response patterns. As supporting evidence, the Cronbach α was calculated for each question pair and listed in Table III. For the five pairs having large number of inconsistent responses (7-8, 11-12, 13-14, 21-22, and 23-24), the reliability coefficients are all smaller than the "acceptable" level (0.65), which again suggest a high level of inconsistency between the measures of the two questions within each pair.

TABLE III. Basic assessment parameters of LCTSR questions (pairs) with college students ($N = 1576$) regarding item mean (or item difficulty), discrimination, and point biserial coefficient (r_{pb}) calculated based on the classical test theory. The Cronbach α of each question pair was also calculated, where values below the acceptable level of 0.65 are marked with * [68].

Question	1	2	3	4	5	6	7	8	9	10	11	12
Mean	0.96	0.95	0.80	0.79	0.64	0.64	0.50	0.70	0.79	0.78	0.51	0.30
Discrimination	0.09	0.11	0.45	0.47	0.72	0.70	0.68	0.54	0.45	0.46	0.62	0.36
r_{pb}	0.22	0.29	0.46	0.49	0.59	0.56	0.48	0.46	0.47	0.47	0.44	0.24
Cronbach α	0.82		0.97		0.89		0.54*		0.92		0.46*	
Question	13	14	15	16	17	18	19	20	21	22	23	24
Mean	0.60	0.51	0.82	0.91	0.83	0.82	0.74	0.68	0.42	0.51	0.44	0.70
Discrimination	0.53	0.49	0.43	0.24	0.37	0.42	0.38	0.50	0.55	0.41	0.42	0.40
r_{pb}	0.37	0.32	0.48	0.42	0.45	0.51	0.31	0.40	0.38	0.26	0.26	0.32
Cronbach α	0.38*		0.69		0.86		0.83		0.56*		0.33*	

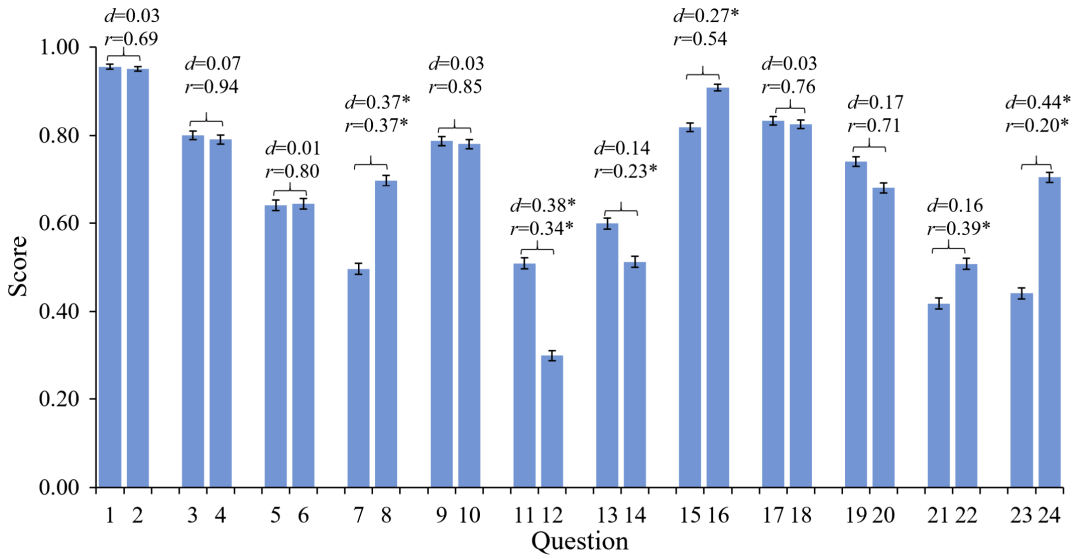

FIG. 2. The average scores of individual questions and the correlation r and Cohen's d between the two individual question scores in each question pair. The values of Cohen's d larger than 0.2 and the values of r less than 0.4 are marked with an * [70].

B. Paired score pattern analysis

As discussed earlier, results of inconsistent responses on questions in two-tier pairs can be used to evaluate the validity of question designs. The percentages of the four responses for each question pair are summarized in Table IV along with the correlations between the question scores of each pair. The results show that the average percentage of the inconsistent patterns for LCTSR is approximately 20% (see the row of "Sum of 01 and 10" in Table IV). Based on the level of inconsistent patterns, two groups of question pairs can be identified ($p < 0.001$, Cohen's $d = 6.4$). Seven pairs (1-2, 3-4, 5-6, 9-10, 15-16, 17-18, and 19-20) appear to have good consistency ($r > 0.5$, $r_{\text{avg}} = 0.76$) as their inconsistent patterns range from 1.7% to 13.7% with an average of 7.4% (SD = 4.6%). The remaining five pairs (7-8, 11-12, 13-14, 21-22, 23-24) have a much higher level of inconsistent responses with an average of 36.4% (SD = 4.4%). The correlation for these is also much smaller ($r < 0.4$, $r_{\text{avg}} = 0.31$).

Interestingly, four of the five pairs with high inconsistency (11-12, 13-14, 21-22, 23-24) are from the pool of new questions added to the LCTSR. Because they were not part of the CTFR-78, these four pairs did not go through the original validity analysis of the 1978 test version. Question pair 7-8 was in the 1978 version of the test but was modified for the multiple-choice format of the LCTSR. This might have introduced design issues specific to its multiple-choice format.

In addition to question design issues, the inconsistent responses can also be the result of guessing and other test-taking strategies, which are more common among low performing students. Therefore, analyzing inconsistent responses against performance levels can help establish the origin of the inconsistent responses. The probability distribution, $P_i(s)$, of inconsistent patterns across different score levels s for all 12 question pairs is plotted in Fig. 4. The error bars are Clopper-Pearson 95% confidence intervals [71,72], which are inversely proportional to the square

TABLE IV. LCTSR paired score patterns (in percentage) of U.S. college freshmen students ($N = 1576$). For the mean values listed in the table, the maximum standard error is 1.12%. Therefore, a difference larger than 3% can be considered statistically significant ($p < 0.05$). The distribution of responses patterns for each question pair is significantly different from random guessing ($p < 0.001$). For example, the results of chi square tests for question pairs 13-14 and 21-22 are $\chi(3) = 240.097$, $p < 0.001$ and $\chi(3) = 55.838$, $p < 0.001$, respectively.

Score patterns	Average	1-2	3-4	5-6	7-8	9-10	11-12	13-14	15-16	17-18	19-20	21-22	23-24
00	22.7	3.3	21.3	31.9	23.9	19.2	42.9	25.3	7.9	13.7	22.6	38.6	21.5
11	57.8	93.8	76.9	58.8	41.7	75.8	23.2	35.7	79.7	79.3	63.7	29.5	35.8
01	11.0	1.2	0.4	4.9	27.7	2.0	7.1	15.8	11.0	2.9	3.4	20.6	34.5
10	8.5	1.7	1.3	4.4	6.7	2.9	26.9	23.3	1.4	4.0	10.3	11.3	8.2
Sum of 01 and 10	19.5	2.9	1.7	9.3	34.4	5.0	33.9	39.1	12.4	6.9	13.7	31.9	42.7
Within pair correlation	0.57	0.69	0.94	0.80	0.37	0.85	0.34	0.23	0.54	0.76	0.71	0.39	0.20

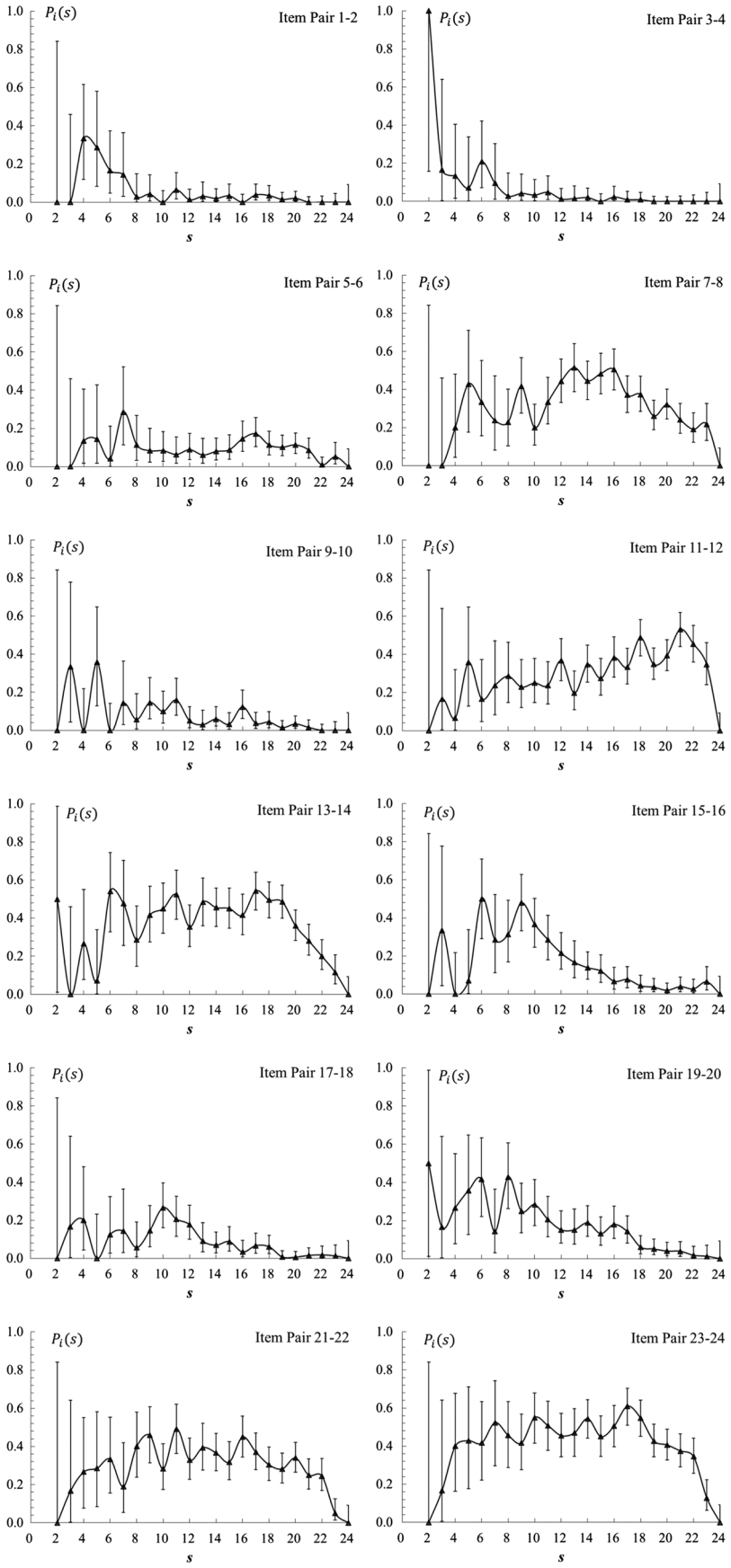

FIG. 3. The distribution of inconsistent patterns as a function of score.

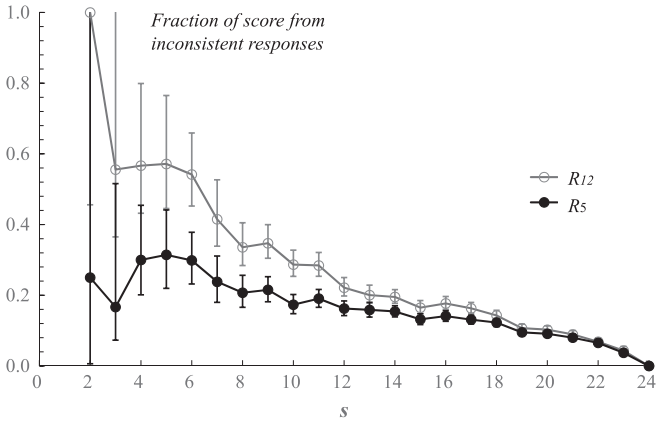

FIG. 4. The fractions of score contributions from all 12 question pairs (R_{12}) and from the 5 inconsistent pairs (R_5) plotted as functions of score.

root of the sample size. At low score levels ($s < 30\%$), the sample sizes are less than 30 [see Fig. 1(b)], which result in large error bars.

The results in Fig. 3 show clear distinctions between the two groups of question pairs with high and low consistency. For the consistent pair group (1-2, 3-4, 5-6, 9-10, 15-16, 17-18, and 19-20), the inconsistent responses come mostly from low performing students; that is, $P_i(s)$ quickly approaches zero when scores increase to the upper 30%. This result suggests that the inconsistent responses for these question pairs are largely due to students guessing, which decreases among high performing students. In contrast, for the inconsistent pair group (7-8, 11-12, 13-14, 21-22, 23-24), the inconsistent responses maintain at high levels throughout most of the score scale and even for students who scored above 90%. Since students with higher scores are expected to answer more consistently than lower performing students [66], the sustained inconsistency at the high score levels is more likely the result of possible question design issues of these five question pairs rather than due to students guessing.

In order to more clearly compare the effects on scores from inconsistent responses at different performance levels, the fractions of score contributions from all 12 questions pairs (R_{12}) and from the 5 inconsistent pairs (R_5) are plotted as functions of score in Fig. 4. The results show that at lower score levels, inconsistent responses are common for all questions ($R_{12} \approx 2R_5$) and are likely the result of guessing. At higher score levels, inconsistent responses come mostly from the 5 inconsistent question pairs ($R_{12} \approx R_5$), and therefore, are more likely caused by issues in question design than by students guessing.

C. The dependence within question pair and between question pairs

As discussed in the methodology section, dependences within and between question pairs were analyzed to study the effects of question designs. The dependences are modeled with the Q_3 fit statistics, which give the correlations of the residuals' variances [Eqs. (1) and (2)]. In this study, a multidimensional Rasch model was used to model the variances from the six subskill dimensions, so that the residuals' variances would primarily contain the contributions from the two-tier structure and other question design features.

For the 24 questions of the LCTSR, there are a total of 276 Q_3 fit statistics, which have a mean value of -0.034 (SD $= 0.130$) for this data set. This mean is within the expected value for a 24-question test $[-1/(24-1) = -0.043]$ and has sufficient local independence at the whole test level to carry out a Rasch analysis. In this study, the main goal of using the Q_3 was to analyze the residual correlations among questions so that the dependences due to question designs could be examined. To do this, the absolute values of Q_3 calculated between every two questions of the test were plotted with a heat map in Fig. 5.

The heat map shows six distinctive groups of questions with moderate to high correlations within each group (>0.2), while the correlations between questions from different subskills are much smaller (<0.2). This result is the outcome of the unique design structure of the LCTSR, which groups questions in order based on the six subskills (see Table I) and the questions within each subskill group also share similar context and design features. For an ideal TTMC test, one would expect to have strong correlations between questions within individual two-tier pairs and much smaller correlations between questions from different two-tier pairs. As shown in Fig. 3, many question pairs show strong correlations between questions within the pairs, however, there are also question pairs, such as Q7-8, in which a question (Q8) is more strongly correlated with questions in other pairs (Q5 and Q6) than with the question in its own two-tier pair (Q7).

In addition, there are six question pairs (11-12, 13-14, 15-16, 21-22, 23-24) for which the within-pair residual correlations are small (<0.2). This suggests that the expected effects of the two-tier structure have been significantly interfered, possibly by other unknown question design features. Therefore, these question pairs should be further examined on their design validities. Among the six question pairs, five are identified as having high inconsistent responses (Table IV). The remaining pair (Q15-16) has much less inconsistency, which also diminished rapidly at high scores (see Fig. 3), and, therefore, will be excused from further inspection.

Synthesizing the quantitative analysis, there appears to be strong indication for design issues for the five question pairs with high inconsistency. Further analysis of these questions using qualitative data will be discussed in the following section to inspect their content validities.

To facilitate the qualitative analysis, the actual answer patterns of the 5 inconsistent question pairs are provided in Table V. Since the total number of possible combinations

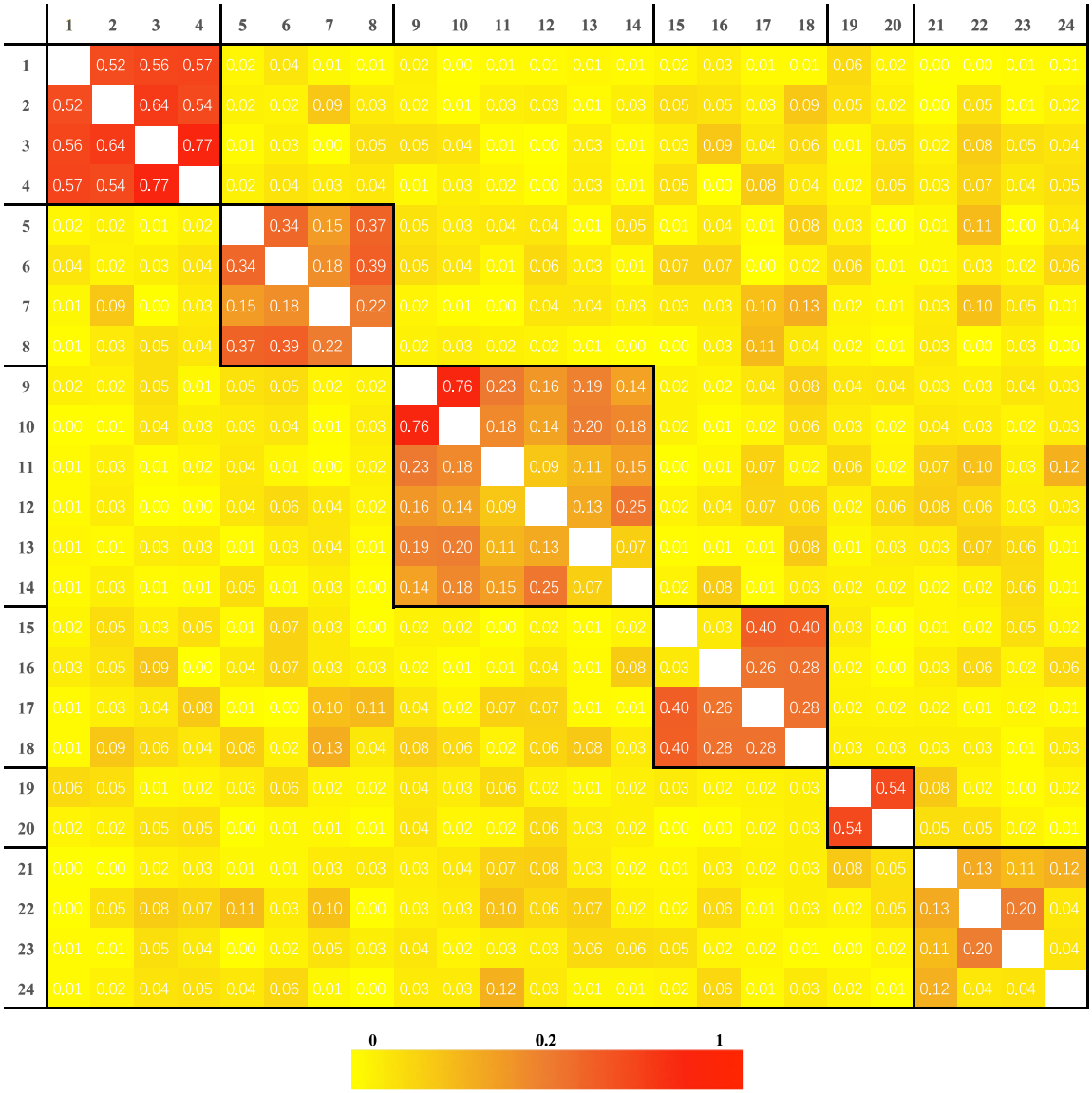

FIG. 5. A heat map of absolute values of Q_3, which gives the residual correlation. The outlined blocks are the six subskills of scientific reasoning (see Table I). The heat map's color transition is nonuniform with the median set at 0.2, which is the recommended Q_3 value for showing a correlation as suggested by Chen and Thissen [64].

TABLE V. The most common answer patterns for questions with high inconsistency. The population is U.S. college freshmen students ($N = 1576$). The numbers reflect the percentage of the corresponding patterns. The patterns are ordered based on their popularity. The correct answer patterns are italicized, which happen to be the most popular answers for all question pairs.

7-8		11-12		13-14		21-22		23-24	
Pattern	%	Pattern	%	Pattern	%	Pattern	%	Pattern	%
da	*41.7*	*ba*	*23.2*	*cd*	*35.7*	*aa*	*29.5*	*ab*	*35.8*
bd	12.8	bb	20.5	ad	12.5	ab	10.2	bb	27.1
aa	10.7	cd	10.9	ce	12.3	eb	7.5	ba	6.1
ba	10.0	ad	10.0	ae	6.7	ea	7.4	cb	6.1
de	5.4	ab	4.4	cc	4.6	bb	6.9	ac	5.6
ea	3.6	de	3.6	bc	4.1	ba	6.7	bc	5.5
ca	3.1	bd	3.5	ac	3.8	ca	5.3	ca	3.6
Others	12.8		24.0		20.3		26.8		10.3

can reach 25, only the most common choices are shown in Table V. For this data set, the most often selected choices (first row of data) happen to be the correct answers. The remaining patterns show a wide variety of combinations of correct and incorrect choices, some of which are used as supporting evidence for student reasoning implied from the qualitative analysis in Sec. IV.

V. RESULTS OF QUALITATIVE VALIDITY EVALUATION

The quantitative analysis suggests that further inspection is needed regarding the question design validity of the five question pairs in LCTSR. To find more concrete evidence regarding the potential question design issues, qualitative studies were conducted, which includes expert evaluations, student written explanations to LCTSR questions, and student interviews. In this section, the qualitative data on the five question pairs will be discussed in detail to shed light on the possible question design issues.

As mentioned previously, a panel of 7 faculty with science and science education backgrounds in physics, chemistry, biology, and mathematics provided expert evaluations of the questions' construct validities. The faculty members were all familiar with the LCTSR test and had used the test in their own teaching and research.

The qualitative data on students' reasoning contains two parts. Part 1 includes open-ended written explanations collected from 181 college freshmen science and engineering majors in a U.S. university. These students were from the same college population whose data are shown in Table II. For this part of the study, students were asked to write down explanations of their reasoning to each of the questions while doing the test. Part 2 of the data includes interviews with a subgroup of the 181 students ($N = 66$). These students were asked to go over the test again and explain their reasoning on how they answered each of the questions.

A. Qualitative analysis of Q7-8

Q7-8 is the second pair of a four-question sequence that measures students' skills on proportional reasoning, which can be accessed from the publicly available LCTSR [19]. The Q7-8 question pair is the second pair in the cluster. The correct answer to Q7 is "d" and the correct reasoning in Q8 is "a." From the results of student scores and response patterns on pair 7-8 as shown in Tables IV and V, almost one-third of students (27.7%) who gave an incorrect answer to Q7, selected the correct reason in Q8. Conversely, 6.7% answered Q7 correctly but provided an incorrect reason in Q8. These results suggest that most students who know the correct outcome also have the correct reasoning, but a substantial number of students who select an incorrect outcome also select the correct reasoning choice. Therefore, the design of the choices might have features contributing to this effect.

The experts' evaluation of LCTSR reported two concerns which might explain these patterns. First, for Q7, choice "d" is the correct answer. However, choice "a" has a value very close to choice "d". As reported by teachers, students are accustomed to rounding numerical answers in pre-college and college assessments. Some students may choose "a" without reading the other choices, thinking that it is close enough for a correct answer. The results in Table V show that 10.7% of the students picked the answer pattern "aa", which gives the correct reasoning in Q8 with a wrong but nearly correct answer of "7½" in Q7.

The second concern came from teachers who encountered students' questions during the test. One common question from students was for clarification about the difference between choices "a" and "e" in Q8. Students often find these to be fundamentally equivalent. Choice "a" is a general statement that the ratio is the same, while choice "e" maintains an identical ratio specifically defined with the actual value. In addition, the wording used in choice "e" of Q8 is equivalent to choice "c" of Q6, which is the correct answer. Since these questions are in a linked set, students may look for consistency among the answer choices, which may lead some students to pick "e" in Q8 instead of "a". From Table V, 5.4% of the students did give the correct answer in Q7 (choice "d") but picked "e" in Q8, which is the most popular "incorrect" explanation for students who answered Q7 correctly.

In order to further explore why students might have selected a correct answer for a wrong reason, and vice versa, students' written explanations and interview responses were analyzed. Figure 6 shows the calculations and written explanations from two students. Student 1 answered Q7 correctly but selected the wrong reasoning

Student 1:

7. 1) Answer: d (correct)

2) Explanation: $\frac{x}{11} = \frac{2}{3} \Rightarrow 3x = 22 \Rightarrow x = \frac{22}{3}$ *or* $7\frac{1}{3}$

8. 1) Answer: e (incorrect, answer key: "a")

2) Explanation: Proportions

Student 2:

7. Answer: b (incorrect, answer key: "d")

Explanation: $\frac{y}{6} = \frac{6}{x} \Rightarrow xy = 36 \Rightarrow x = 9$, *just use ratios*

8. Answer: a (correct)

Explanation: use ratios

FIG. 6. Examples of two students' written explanations to Q7-8. Student 1 had the correct ratio of 2:3 from the previous question pair Q5-6 and applied the proportional reasoning correctly in Q7. However, the student selected "e" in Q8 as the reasoning, which is incorrect. Student 2 knew to use the ratio, but failed to apply it to obtain the correct result.

in Q8. The student appears to have a good understanding of how to use the mathematics involved in proportions and ratios, but picked "e" as the answer for the reasoning part. Among the 66 interviewed, 5 students had answers and explanations similar to that of student 1 shown in Fig. 6. These students could correctly solve the problem using knowledge of ratios and mathematics for Q7, but picked "e" for their reasoning. Further analysis of students' explanations in the interviews revealed that they all thought "e" was a more detailed description of the ratio, which is specific to the question, whereas "a" was just a generally true statement. Therefore, these students chose the more descriptive answer out of the two choices that both appeared to them to be correct.

In the example of student 2 shown in Fig. 6, the student failed to set up and solve the proportional relations in the question, but picked the correct reason. The explanation was rather simplistic and simply indicated that the question involved a ratio. Apparently, the student was not able to correctly apply the ratio to solve this problem, but knew to "use ratios", which led to correctly selecting "a" in Q8.

These results indicate that the design of choices "a" and "d" in Q8 is less than optimal. Students may treat the two choices as equivalent with choice "a" being a generally true statement about ratio and choice "e" being a more substantive form of the same ratio. This design can cause confusion to those who are competent in handling proportions, but is more forgiving to those who are less competent, leading to both false positives and false negatives [57]. In addition, although it was not observed in our interviews, the numerical value of the incorrect choice "a" ($7\ 1/2$) was very close to the correct answer (choice "d", $7\ 1/3$), which might cause false negatives due to rounding, as suggested by our experts. Therefore, it would be beneficial to modify the numerical value of choice "a" such that it would not be in the range of typical rounding of the correct answer.

B. Qualitative analysis of Q11-12 and Q13-14

These two question pairs were designed to measure student reasoning on control of variables (COV) using a context that involves experiments with fruit flies. The correct answers to the four individual questions are "b," "a," "c," and "d". From Table II, the assessment results of Q11-12 and Q13-14 show a high level of inconsistent answer patterns (01 and 10).

For these two pairs of questions, the expert review reported four areas of concerns. First, the graphical representation of black paper wrapped on the test tube is misleading. Several faculty pointed out that the black dots on the paper were initially perceived as flies, rather than black paper, and they had to go back and forth to read the text and figure in order to make sense of the scenario. This may not be a critical validity issue but does impact information processing regarding time and accuracy, and may lead to confusion and frustration among students. Another related issue is the presentation of the numbers of flies in the different areas of the tube. For tubes without the black paper, the numbers are straightforward. For tubes with black paper, the numbers of flies in the covered part of the tube are not shown. This may not be an issue for advanced students but can pose some processing overhead for some students, particularly those in lower grades. For consistency of representation, it is suggested that the number of flies in the covered part of the tube also be shown on the figure.

Second, the orientation of the tubes relative to the provided red and blue light may also cause confusion. There is no indication of a table or other means of supporting the tubes, and the light is shown as coming from both top-down and bottom-up directions on the page. In a typical lab setting, light coming from bottom-up will be blocked by a table top (usually made of non-transparent materials). The current configuration is not common in everyday settings and may mislead students to consider that all tubes are lying flat (horizontally) on a table surface (overhead view) with the light beams projected horizontally from opposite directions. In this interpretation, it will be difficult to understand how gravity plays a role in the experiment.

The third concern is the design of choices for the reasoning questions (Q12 and Q14). The typical scientific strategies for analyzing data from a controlled experiment is to make explicit comparisons between the results of different conditions to emphasize the covariation relations established with controlled and varied variables. In this set of questions, the correct comparisons should include tubes III and IV to determine the effect of gravity and tubes II and IV to determine the effect of light. Tube IV is the base for comparison, with which both the gravity and light are set in a noneffect state. However, none of the choices in Q12 and Q14 gives a straightforward statement of such comparisons and tube IV is not included in the correct answers (choice "a" for Q12 and choice "d" for Q14). The language used in the choices simply describes certain factual features of the setting but doesn't provide clear comparisons.

In addition, tube I is a confounded (uncontrolled) situation in which both the gravity and light are varied with respect to tube IV. Correct COV reasoning should exclude such confounded cases but tube I is included in the correct answer of Q14 (choice "d"), which raises concerns about the validity of this choice being correct.

Finally, it is worth noting that the conditions in the questions do not represent a complete set. For example, the configurations for the absence of light or gravity are not included in the settings. Therefore, the outcomes should not be generalized into those situations.

Results from students' written explanations and interviews support the experts' concerns. Many students complained about the images in Q11-Q14. Some students were

confused because "*the black paper looks lumpy and looks like a mass of flies*". Others asked if the black paper blocked out the light entirely as the light may come in from the ends as illustrated by the rays in the figure.

Students were also confused on how the tubes were arranged. Some thought that all the tubes were lying on a table top and the figures were a bird's eye view of the setup. It was only after the interviewer sketched a table perspective overlaid with the tubes that many realized how the tubes were meant to be positioned. In addition, the arrows of the light seemed to cause misinterpretation of the meaning of "respond to light." Some students thought that the light beams were only going in the two directions illustrated by the arrows: "*the flies fly towards the light (against the light beam) in tube I and III*". This can cause complications in understanding the effect of gravity and the outcomes of the horizontal tubes.

There are also students who seemed to have a different interpretation of what "respond to gravity" means. For example, for students who selected "a" (red light not gravity) in Q11, their typical reasons included: "*if the flies responded to gravity, they should go to the bottom of tube I and III*," or if "*the flies ignore gravity, they fly high in tubes*." This result suggests a possible complication in reasoning with gravity. Since gravity in real life makes things fall downward, being able to defy gravity (flying up) seemed to have been made equivalent to the noneffect of gravity by these students.

In general, students did not like the choices for the reasoning provided in Q12 and Q14. It appeared that students were looking for choices that made direct comparisons of multiple tubes, which is the correct strategy in analyzing data of COV experiments. Unsatisfied with the choices, some students also suggested their own answers; for example: "*The files are even in IV, but are concentrated in the upper end of III, and nearly even in II*"; and "*Comparison of III and IV shows response to gravity while comparison of II and IV shows response to light*."

A few students also expressed concern about the interaction between light and gravity: "The question cannot tell how the flies react to light and gravity separately and one may need BOTH light and gravity to get a reaction, because the light is never turned off."

The response patterns in Table III show that for Q11-12 the major inconsistent pattern is "bb" (20.5%), which is correct for Q11 but incorrect for Q12. A typical student answer (student 3) is shown in Fig. 7. The student's reasoning was mostly based on the flies' behaviors in responding to gravity and was implicit regarding the effect of red light. Students might implicitly know that the red light was not a factor, and then primarily focused on the influential factor (the gravity) in their explanations. This also agrees with many two-tier assessment studies which have demonstrated that knowing is easier than reasoning [34,41].

Student 3:
11. Answer: b (correct).
Explanation: They move up regardless of light.
12. Answer: b (incorrect, answer key: "a")
Explanation: Most flies did not go to the bottom of tubes I and III, it represents the results.

Student 5:
13. Answer: c (correct)
Explanation: Comparing II and IV, they took preference of light when gravity was a none factor.
14. Answer: e (incorrect, answer key: "d")
Explanation: Initially thought d, but e was a better answer. You can see they respond to gravity but in Tube II & IV, you can tell blue light is also a responsive factor. Options d and e are confusingly similar. Did not like the answer choices.

Students' Comments on Q11-14:
- *The black paper on tubes I and II looks like a mass of flies.*
- *Maybe pair III&I and II&IV to compare in the answers.*
- *Q14 -- options d and e are confusingly similar.*
- *The answer choices were not good, does not seem like good choices.*
- *None of the answers seem correct.*

FIG. 7. Examples of students' written explanations to Q11-14.

For Q13-14, many students seemed to struggle between choices "d" and "e." A significant number of students, who answered correctly for Q13, picked the incorrect choice "e" for Q14. A representative student response (student 5) in Fig. 7 shows that the student used correct reasoning in comparing the tubes but was not satisfied with the answers in Q14 and picked the wrong choice. This indicates that the design of the choices of Q14 can benefit from further refinement. In particular, the correct answer ("d") should not include the confounded condition (tube I).

From student interviews and response patterns, there is also evidence that suggests that students' understanding of statistical significance of a measured outcome is an important factor in their decision making regarding the covariation effect of an experiment. For example, approximately 15% of the students considered red light as an influential factor in Q11-12. Their reasoning indicates an issue in understanding statistical significance: "*it is hard to decide because in tube II, 11 flies are in the lighted part and 9 in the dark part, a* 11:9 *ration, which is not a very big difference to make judgement*." It appears that these students were not sure if 11 means more flies or just a random uncertainty of an equal distribution. To reduce such confusion, it might be helpful to increase the total number of flies so that students are better able to make a judgement between random and meaningful outcomes.

The results suggest that the pictorial presentations of the questions did have some negative impact on student's understanding of the actual conditions of the experiments. In addition, the design of the choices in the reasoning questions are less than optimal. In COV conditions, students have the natural tendency to compare tubes rather than look at a single tube to test a hypothesis. Therefore, the choices should reflect such comparisons and the confounded cases should not be included in the correct answer.

C. Qualitative analysis of Q21-22 and Q23-24

The last two pairs of the LCTSR were designed to measure student ability in hypothetical deductive reasoning [73,74]. As discussed earlier, these two pairs of questions have a different style than the traditional two-tier design. Q21-22 asks for the experimental design to test a given hypothesis in tier 1 and the experimental outcome that will disprove the hypothesis in tier 2. Meanwhile, in Q23-24, a phenomenon is introduced along with two possible hypotheses and an experimental design that can test the hypotheses. Students are asked to select the experimental outcomes that will disprove the two hypotheses in tier 1 and tier 2, respectively.

The expert reviews reported a number of concerns with these two question pairs. Both involve complex experimental settings with a large number of detailed variables and possible extensions. Subsequently, they require a high level of reading and processing that can be overly demanding for some students. In terms of the content and reasoning, the experts cited a number of plausibility and unclear variable issues. For Q21-22, the scenario involves additional variables and relations including dry ice, a balloon, and suction, which are not clearly defined or discussed for their roles and uses in the experiment. Students may interpret some of these relations in a way tangential to what the questions were designed for, which would render their reasoning irrelevant.

More specifically, the concept of suction and the use of a balloon may cause additional confusion. Suction is a physical scenario in which the air enclosed in a sealed volume is mechanically removed to create a lower pressure region relative to the exterior environment, causing liquid to be pushed into the lower pressure region by the higher air pressure in the environment. For Q21-22, although the dissolution of carbon dioxide will remove a certain amount of gas in the upside-down glass and decrease the pressure within, the process is mechanically different from a real-world suction event in which the gas is removed through mechanical means. Such differences may confuse students and cause unintended interpretations. Even when the measurement of pressure is granted in this experiment, from the practical standpoint, using a balloon should be avoided. Based on life experience, a typical balloon is made of rubber material and will melt or burn if placed on top of a candle flame. This setting is not plausible to function as proposed by the questions.

For Q23-24, the experts again identified plausibility issues. Most plastic bags are air and water tight and do not work like osmosis membranes. Furthermore, liquid is nearly incompressible, and solutions of positive or negative ions will not produce nearly enough pressure needed to compress liquid (if at all). Therefore, the proposed scenario of compressing liquid-filled plastic bags is physically implausible and the proposed experiment may leave students with a sense of unreality. In this case, the experimental setting mixes real and imaginary parts of common sense knowledge and reasoning. In addition, it asks students to remove certain known features of real objects and retain other features to form a causal relation. This is an awkward design, which requires one to partially shut down common sense logic while pretending to use the same set of common sense logic to operate normally and ignore certain common-sense outcomes. This is highly arguable to be an appropriate context for students to perform hypothetical reasoning, which is already complex itself.

In addition, the two pairs of questions ask for proving a hypothesis wrong. Typically, students have more experience in supporting claims as correct rather than as wrong. Therefore, students may answer incorrectly due to simple oversight and or misreading the question. Bolding or underlining key words may alleviate these issues.

Students' responses to these questions resonate with the experts' concerns. From interviews and written explanations, many students complained that these questions were difficult to understand and needed to be read multiple times. Many continued to misinterpret the questions which impacted their responses. For example, for the task which asks students to prove the hypothesis wrong, some students thought it meant "*what would happen if the experiment went wrong?*" Others simply went on to prove the hypothesis was right. This is evident from the large number of "ab" patterns (10.2%) on Q21-22 and "bb" patterns (27.1%) on Q23-24. In addition, since questions Q23-24 ask for proving two hypotheses, this cued some students to compare between the two and left them struggling with the relations: "*If explanation I is wrong, does it mean that explanation II has to be right?*" and "*I misread the question the 1st time, I didn't see that they were 2 separate explanations; thought one would be right and the other would be wrong*." This misunderstanding of the question caused some students to prove one hypothesis wrong and one correct.

In addition, for Q21-22, students overwhelmingly disliked the inclusion of the concept of suction and a balloon in the scenario: "*suction seems unrelated to the possible explanation;*" "*suction has nothing to do with what is happening;*" "*choice d was unreasonable;*" "*the question made no mention of a balloon being available for use—option d was wrong. The starting part of option D was wrong because the use of the word 'suction' was wrong, because pressure was what caused the rise*."

There were also students not comfortable with the use of dry ice: "did not understand the process described in the question—thought dry ice was to frozen the water;" "for this question it seems like you need to know chemistry;" "concerned about answer A because is there a guarantee that the carbon dioxide isn't going to escape the water? How long will it reside in the water?"

Based on students' responses and expert reviews, the design of these two question pairs appears to be less than optimal. From the expert point of view, there are significant

plausibility concerns to some of the involved conditions. Although such concerns were not observed from students' responses, these should be addressed in order to make the science accurate. From the students' perspective, these questions require substantial content knowledge in chemistry and physics in order to fully understand the intentions of the experimental designs and processes. A lack of content understanding, such as that involved with dry ice, solubility of carbon dioxide, osmosis processes, etc., could interfere with students' ability to reason out solutions to the questions. On the other hand, students fluent in such content knowledge may simply skip the intended reasoning process and use memorized scientific facts to answer the questions. Therefore, it would be helpful to address the identified content knowledge issues so that students' hypothetical deductive reasoning abilities can be more accurately assessed. And last, the questions asked students to prove a provided hypothesis wrong, which led to significant misreading among students. This type of design can also benefit from further revisions.

D. Ceiling effect

In general, the Lawson test is known to be a simple test that measures learners' basic reasoning skills, and the ceiling effect is evident when it is used with students at the college level. In order to understand the usability of the Lawson's test for different age groups, Bao *et al.* [4] reported on the general developmental trend of Lawson test scores for Chinese and U.S. students. They showed that the LCTSR scores approach ceiling for age groups around the senior high school to early college years.

One concern of the demonstrated ceiling effect is that it is significantly below the 100% mark. When including senior college students, the maximum average is approximately 80%. One would expect the ceiling to be close to 100% for advanced students. Based on the analysis of students' qualitative responses, this low ceiling level is likely due to the question design issues discussed in this paper. That is, it was evident that some of the senior college students were able to apply appropriate reasoning to obtain correct answers in tier 1, but were in disagreement with the choices of reasoning in tier-2, such as those discussed for Q8, Q12, and Q14. The ceiling effect also limits the discrimination of LCTSR among students beyond high school. Therefore, in order to assess the reasoning abilities of college students, additional questions that involve more advanced reasoning skills are needed.

VI. CONCLUSION AND DISCUSSION

In general, the LCTSR is a practical tool for assessing large numbers of students. The instrument has good overall reliabilities with both the individual and pair scoring methods (Cronbach $\alpha > 0.8$) when the test length is controlled. The test is on the easy side for college students with typical mean scores at the 60% level for pair scoring and 70% level for individual scoring. The residual correlation analysis using Q_3 statistics shows acceptable local independence for the overall test.

When individual question pairs are examined, however, seven out of the twelve pairs perform as expected for the TTMC design, while the remaining five pairs suffer from a variety of design issues. These five question pairs have a considerable amount of inconsistent response patterns. Qualitative analysis also indicates wide ranging question design issues among the content and context scenarios, pictorial and narrative presentations, and answer choices. All of these potentially contribute to the high level of inconsistent responses and the less than optimal ceiling.

The identified question design issues are likely an underlying cause for the substantial uncertainties in students' responses, which need to be carefully considered when interpreting the assessment results. While the assessment of the overall scientific reasoning ability of large classes is reliable, as the uncertainties from the question design issues can be "averaged out," interpretations of the affected subskills including *proportional reasoning, control of variables,* and *hypothetic-deductive reasoning*, may have less power due to the significant amount of uncertainty involved. Direct comparisons among the affected subskills without addressing the question design issues is questionable. For example, a recent study has shown that the progression trends of COV and HD measured with subskill scores are different from that of the other subskills [38]. Their results were used to conclude a distinctive stage of scientific reasoning, which should be further inspected against the question design issues uncovered in this study. Therefore, based on the results of this study, it is suggested that the LCTSR is generally valid for assessing a unidimensional construct of scientific reasoning; however, its validity in assessing subskills is limited.

In response to the limitations of the LCTSR observed in practice, several studies have attempted to design new questions for assessing scientific reasoning [75,76]. For example, Zhou *et al.* [76] developed questions specific for COV. These development efforts can be further informed by the results from this study, which provides a comprehensive and concrete evaluation of the possible question design issues of the LCTSR.

The method used in this study, which emphasizes the inconsistent responses of question pairs, may also contribute to the methodology in developing and validating two-tier tests. While the two-tier test was originally developed as an effective alternative of the traditional multiple-choice test to suppress the possible false positive [34,42,57], it has been unclear as to how to quantify the likelihood of false positives in TTMC tests. The method of analyzing inconsistent response patterns described in this paper provides a practical means to quantitatively evaluate the probabilities

of possible false positives and negatives, which can be used as evidence for establishing the validity of two-tier tests.

Finally, this study conducted validity research on the LCTSR using only the data from U.S. college freshman students. The validity evaluation is population dependent and additional validity issues may exist among other populations. Nevertheless, the claim regarding poor content validity for the five question pairs in LCTSR is sufficiently warranted with the evidence identified.

ACKNOWLEDGMENTS

The research is supported in part by the National Science Foundation Grants No. DUE-1044724, DUE-1431908, DRL-1417983, and DUE-1712238. 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, *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century* (National Academies Press, Washington, DC, 2012).

[2] J. Piaget, The stages of the intellectual development of the child, in *Readings in Child Development and Personality*, edited by P. H. Mussen, J. J. Conger, and J. Kagan (Harper and Row, New York, 1970), pp. 98–106.

[3] J. Hawkins and R. D. Pea, Tools for bridging the cultures of everyday and scientific thinking, J. Res. Sci. Teach. **24**, 291 (1987).

[4] 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).

[5] C. Zimmerman, The development of scientific reasoning: What psychologists contribute to an understanding of elementary science learning, Paper commissioned by the National Academies of Science (National Research Council's Board of Science Education, Consensus Study on Learning Science, Kindergarten through Eighth Grade) (2005).

[6] C. Zimmerman, The development of scientific thinking skills in elementary and middle school, Dev. Rev. **27**, 172 (2007).

[7] P. Adey and M. Shayer, *Really Raising Standards* (Routledge, London, 1994).

[8] Z. Chen and D. Klahr, All other things being equal: Acquisition and transfer of the control of variables strategy, Child Development **70**, 1098 (1999).

[9] A. M. Cavallo, M. Rozman, J. Blickenstaff, and N. Walker, Learning, reasoning, motivation, and epistemological beliefs, J. Coll. Sci. Teach. **33**, 18 (2003).

[10] M. A. Johnson and A. E. Lawson, What are the relative effects of reasoning ability and prior knowledge on biology achievement in expository and inquiry classes?, J. Res. Sci. Teach. **35**, 89 (1998).

[11] V. P. Coletta and J. A. Phillips, Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability, Am. J. Phys. **73**, 1172 (2005).

[12] H.-C. She and Y.-W. Liao, Bridging scientific reasoning and conceptual change through adaptive web-based learning, J. Res. Sci. Teach. **47**, 91 (2010).

[13] S. Ates and E. Cataloglu, The effects of students' reasoning abilities on conceptual understandings and problem-solving skills in introductory mechanics, Eur. J. Phys. **28**, 1161 (2007).

[14] J. L. Jensen and A. Lawson, Effects of collaborative group composition and inquiry instruction on reasoning gains and achievement in undergraduate biology, CBE-Life Sci. Educ. **10**, 64 (2011).

[15] B. Inhelder and J. Piaget, *The Growth of Logical Thinking from Childhood to Adolescence: An Essay on the Construction of Formal Operational Structures* (Basic Books, New York, 1958).

[16] A. E. Lawson, The development and validation of a classroom test of formal reasoning, J. Res. Sci. Teach. **15**, 11 (1978).

[17] V. Roadrangka, R. H. Yeany, and M. J. Padilla, *GALT, Group Test of Logical Thinking* (University of Georgia, Athens, GA, 1982).

[18] K. Tobin and W. Capie, The development and validation of a group test of logical thinking, Educ. Psychol. Meas. **41**, 413 (1981).

[19] A. E. Lawson, Lawson classroom test of scientific reasoning. Retrieved from http://www.public.asu.edu/~anton1/AssessArticles/Assessments/Mathematics%20Assessments/Scientific%20Reasoning%20Test.pdf (2000).

[20] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, Phys. Teach. **30**, 141 (1992).

[21] C. Pratt and R. G. Hacker, Is Lawson's classroom test of formal reasoning valid?, Educ. Psychol. Meas. **44**, 441 (1984).

[22] G. P. Stefanich, R. D. Unruh, B. Perry, and G. Phillips, Convergent validity of group tests of cognitive development, J. Res. Sci. Teach. **20**, 557 (1983).

[23] G. M. Burney, The construction and validation of an objective formal reasoning instrument, Ph.D. Dissertation, University of Northern Colorado, 1974.

[24] F. Longeot, Analyse statistique de trois tests genetiques collectifs, Bull. Inst. Natl. Etude **20**, 219 (1965).

[25] R. P. Tisher and L. G. Dale, *Understanding in Science Test* (Australian Council for Educational Research, Victoria, 1975).

[26] K. Tobin and W. Capie, The test of logical thinking: Development and applications, *Proceedings of the Annual Meeting of the National Association for Research in Science Teaching, Boston, MA* (1980).

[27] J. A. Rowell and P. J. Hoffmann, Group tests for distinguishing formal from concrete thinkers, J. Res. Sci. Teach. **12**, 157 (1975).
[28] M. Shayer and D. Wharry, *Piaget in the Classroom Part 1: Testing a Whole Class at the Same Time* (Chelsa College, University of London, 1975).
[29] R. G. Hacker, The construct validities of some group tests of intellectual development, Educ. Psychol. Meas. **49**, 269 (1989).
[30] M. D. Reckase, T. A. Ackerman, and J. E. Carlson, Building a unidimensional test using multidimensional items, J. Educ. Measure. **25**, 193 (1988).
[31] A. Hofstein and V. N. Lunetta, The laboratory in science education: Foundations for the twenty-first century, Sci. Educ. **88**, 28 (2004).
[32] V. P. Coletta, J. A. Phillips, and J. J. Steinert, Interpreting force concept inventory scores: Normalized gain and SAT scores, Phys. Rev. ST Phys. Educ. Res. **3**, 010106 (2007).
[33] C.-Q. Lee and H.-C. She, Facilitating students' conceptual change and scientific reasoning involving the unit of combustion, Res. Sci. Educ. **40**, 479 (2010).
[34] Y. Xiao, J. Han, K. Koenig, J. Xiong, and L. Bao, Multilevel Rasch modeling of two-tier multiple choice test: A case study using Lawson's classroom test of scientific reasoning, Phys. Rev. Phys. Educ. Res. **14**, 020104 (2018).
[35] K. Diff and N. Tache, From FCI to CSEM to Lawson test: A report on data collected at a community college, AIP Conf. Proc. **951**, 85 (2007).
[36] V. P. Coletta, Reaching more students through thinking in physics, Phys. Teach. **55**, 100 (2017).
[37] S. Isaac and W. B. Michael, *Handbook in Research and Evaluation: A Collection of Principles, Methods, and Strategies Useful in the Planning, Design, and Evaluation of Studies in Education and the Behavioral Sciences* (Edits publishers, San Diego, 1995).
[38] 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).
[39] L. Ding, X. Wei, and X. Liu, Variations in university students' scientific reasoning skills across majors, years, and types of institutions, Res. Sci. Educ. **46**, 613 (2016).
[40] L. Ding, X. Wei, and K. Mollohan, Does higher education improve student scientific reasoning skills?, Int. J. Sci. Math. Educ. **14**, 619 (2016).
[41] 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).
[42] D. F. Treagust, Development and use of diagnostic tests to evaluate students' misconceptions in science, Int. J. Sci. Educ. **10**, 159 (1988).
[43] C.-C. Tsai and C. Chou, Diagnosing students' alternative conceptions in science, J. Comput. Assist. Learn. **18**, 157 (2002).
[44] 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).
[45] 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).
[46] 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).
[47] H.-S. Lee, O. L. Liu, and M. C. Linn, Validating measurement of knowledge integration in science using multiple-choice and explanation items, Appl. Meas. Educ. **24**, 115 (2011).
[48] 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).
[49] D. K.-C. Tan and D. F. Treagust, Evaluating students' understanding of chemical bonding, Sch. Sci. Rev. **81**, 75 (1999); https://repository.nie.edu.sg/handle/10497/14150.
[50] 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).
[51] 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).
[52] J.-R. Wang, Evaluating secondary students' scientific reasoning in genetics using a two-tier diagnostic instrument, Int. J. Sci. Math. Educ. **2**, 131 (2004).
[53] 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).
[54] 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).
[55] J. Yao and Y. Guo, Validity evidence for a learning progression of scientific explanation, J. Res. Sci. Teach. **55**, 299 (2017).
[56] J. Yao, Y. Guo, and K. Neumann, Towards a hypothetical learning progression of scientific explanation, Asia-Pac. Sci. Educ. **2**, 4 (2016).
[57] D. Hestenes and I. Halloun, Interpreting the force concept inventory: A response to Huffman and Heller, Phys. Teach. **33**, 502 (1995).
[58] J. Wang and L. Bao, Analyzing force concept inventory with item response theory, Am. J. Phys. **78**, 1064 (2010).
[59] J. W. Creswell and J. D. Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (Sage Publications, Thousand Oaks, CA, 2011).
[60] H. Wainer and G. L. Kiely, Item clusters and computerized adaptive testing: A case for testlets, J. Educ. Measure. **24**, 185 (1987).
[61] P. R. Rosenbaum, Items bundles, Psychometrika **53**, 349 (1988).
[62] 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.

[63] W. M. Yen, Effects of local item dependence on the fit and equating performance of the three-parameter logistic model, Appl. Psychol. Meas. **8**, 125 (1984).

[64] W.-H. Chen and D. Thissen, Local dependence indexes for item pairs using item response theory, J. Educ. Behav. Stat. **22**, 265 (1997).

[65] W. M. Yen, Scaling performance assessments: Strategies for managing local item dependence, Educ. Meas. **30**, 187 (1993).

[66] A. W. Gotwals and N. B. Songer, Reasoning up and down a food chain: Using an assessment framework to investigate students' middle knowledge, Sci. Educ. **94**, 259 (2010).

[67] J. Yasuda, N. Mae, M. M. Hull, and M. Taniguchi, Analyzing false positives of four questions in the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010112 (2018).

[68] R. F. DeVellis, *Scale Development: Theory and Applications* (Sage Publications, Thousand Oaks, CA, 2012).

[69] L. Ding, R. Chabay, B. Sherwood, and R. Beichner, Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment, Phys. Rev. ST Phys. Educ. Res. **2**, 010105 (2006).

[70] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Lawrence Earlbaum, Hillsdale NJ, 1988), p. 20.

[71] C. J. Clopper and E. S. Pearson, The use of confidence or fiducial limits illustrated in the case of the binomial, Biometrika **26**, 404 (1934).

[72] J. T. Willse, Classical test theory functions (R package version 2.3.2). Retrieved from https://CRAN.R-project.org/package=CTT (2018).

[73] A. E. Lawson, The generality of hypothetic-deductive reasoning: Making scientific thinking explicit, Am. Biol. Teach. **62**, 482 (2000).

[74] A. E. Lawson, B. Clark, E. Cramer-Meldrum, K. A. Falconer, J. M. Sequist, and Y.-J. Kwon, Development of scientific reasoning in college biology: Do two levels of general hypothesis-testing skills exist?, J. Res. Sci. Teach. **37**, 81 (2000).

[75] J. Han, Scientific reasoning: Research, development, and assessment, Ph.D. thesis, The Ohio State University, 2013.

[76] S. Zhou, J. Han, K. Koenig, A. Raplinger, Y. Pi, D. Li, H. Xiao, Z. Fu, and L. Bao, Assessment of scientific reasoning: The effects of task context, data, and design on student reasoning in control of variables, Think. Ski. Creat. **19**, 175 (2016).