



Abstract: *The Lawson's Classroom Test of Scientific Reasoning (LCTSR) is a popular instrument that measures the development of students' scientific reasoning skills. The instrument has a two-tier question design, which has led to multiple ways of scoring and interpretation. In this research, a method of pattern analysis was proposed and applied to analyze two-tier item pairs on the subskill of Control-of-Variable (COV) of LCTSR. The data were collected from students in grade 4 through college in both the United States and China. Students' response patterns to two combined item pairs were analyzed and compared against students at different grade levels and reasoning development levels. Six performance levels were established based on students' response patterns, serving as indicators of COV reasoning development levels. With the new method, a relation was obtained between students' level of COV skills and grade level, as well as their level of overall reasoning development. It can provide useful information on the possible developmental levels of students' reasoning skills.*

Keywords: *control-of-variable, grade level, pattern analysis, scientific reasoning*

Shao-Na Zhou

South China Normal University, China

Qiao-Yi Liu

The Ohio State University, USA

Kathleen Koenig

University of Cincinnati, USA

Qiu-ye Li-Yang Xiao

South China Normal University, China

Lei Bao

The Ohio State University, USA

ANALYSIS OF TWO-TIER QUESTION SCORING METHODS: A CASE STUDY ON THE LAWSON'S CLASSROOM TEST OF SCIENTIFIC REASONING

**Shao-Na Zhou,
Qiao-Yi Liu,
Kathleen Koenig,
Qiu-ye Li-Yang Xiao,
Lei Bao**

Introduction

In science, technology, engineering, and mathematics (STEM) education, scientific reasoning has gained increasing attention recently from science educators and researchers. In general, the scope of scientific reasoning encompasses various thinking and reasoning skills involved in inquiry, experimentation, evidence evaluation, inference, and argumentation that support the formation and modification of concepts and theories about the natural and social world (Zimmerman, 2007). It plays a crucial role in the development of students' creative thinking and problem-solving skills, which are necessary when managing real-world situations in professions beyond the classroom (iSTAR Assessment, 2010; Zhou et al., 2016). In K-12 education, it has been reported that effort in improving scientific reasoning skills has a profound impact on students' academic performance (Adey & Shayer, 1990), and a positive correlation has been reported between students' scientific reasoning abilities and course achievement (Coletta & Phillips, 2005). Furthermore, it has been shown that the level of students' scientific reasoning skills is a better indicator of their success in biology learning, comparing with their prior knowledge (Johnson & Lawson, 1998). These results prompt the need for students to develop a strong set of scientific reasoning skills, alongside a solid foundation of content knowledge.

A number of studies have been conducted to determine the mechanism of acquiring scientific reasoning skills, as well as effective teaching strategies that aid students in learning, retaining, and transferring these skills (Dunbar & Klahr, 2012; Joep, et al., 2019). In a review on scientific reasoning, Zimmerman claimed that investigation skills can interact with science knowledge to create a relationship that promotes scientific reasoning development (Zimmerman, 2007). It has also been reported that scientific reasoning requires a complex set of cognitive skills, the development of which follows a prolonged path, and students' performance varies as they progress along this path (Zimmerman, 2007).

In order to quantitatively measure the development of students' scientific reasoning skills, several assessment instruments have been developed and employed by education researchers, such as the group assessment of logical thinking test (GALT), the test of logical thinking (TOLT), and the Lawson's Classroom Test of Scientific Reasoning (LCTSR) (Lawson, 1978). The LCTSR is a popular assessment instrument that investigates students' scientific reasoning skills from primary school to university (Bao et al., 2009), and the validity of its current version has been analyzed in detail (Xiao et al., 2018). Nevertheless, multiple scoring methods of the two-tier LCTSR test have been proposed, and generated different interpretations (Xiao et al., 2018). More specifically, two-tier multiple-choice item pairs with different difficulties are generally not distinguished. However, this dimension of difficulty can provide insights regarding students' scientific reasoning progress. This work details a cross-sectional study using data from students in both the United States and China on the correlation between students' answer patterns and their level of scientific reasoning for different grade levels and difficulty levels, which further extends the application of LCTSR on measuring the development of students' scientific reasoning skills.

Literature Review on Lawson's Classroom Test of Scientific Reasoning (LCTSR)

The development of LCTSR

Historically, the Piagetian tasks were considered to be a standard method of measuring students' scientific reasoning skills, which, however, are time-consuming and required experienced interviewers, special materials, and equipment (Goldschmid, 1967; Lawson & Blake, 1976; Lawson et al., 1975). In 1978, Lawson designed an assessment instrument that measures students' level of scientific reasoning development, called the Lawson's classroom test of formal reasoning (CTFR-78). The paper and pencil style of the CTFR-78 addressed the need for a reliable, convenient assessment tool that would be more practical for classroom use, compared to the Piagetian tasks.

A paper and pencil test, compared to clinical interview tasks, not only requires the ability to read and write, but also provides little motivation for the test takers from the materials or equipment being used, since it is not as personal or relaxed. Taking these challenges into consideration, Lawson aimed to strike a balance between the convenience of paper and pencil tests and the interaction of interview tasks in CTFR-78. CTFR-78 involves an instructor performing a demonstration in front of a class, after which the instructor would pose a question to the entire class, and the students would mark their answers in their test booklets. The booklets contain the questions followed by several answer choices. For each of the test items, students had to choose the correct answer and provide a reasonable explanation in order to receive credit for that item, forming the two-tier test design. In order to establish the validity of the test, Lawson administered CTFR-78 to 513 students from 8th through 10th grade, and selected 72 of them to participate in clinical interviews with Piagetian tasks that reflected the three established levels of reasoning (concrete reasoning, transitional reasoning, and formal reasoning) (Lawson, 1978). After comparing the test scores with students' response to interview tasks, Lawson found that the results from CTFR-78 and the clinical interviews had a good agreement, while CRFR-78 might have a tendency to underestimate students' scientific reasoning ability slightly. The validity of CTFR-78 was further established by other researchers (Pratt & Hacker, 1984; Stefanich et al., 1983), with item analysis and principle-components analysis.

In 2000, building on previous work, Lawson developed an improved version of the assessment instrument, named Lawson's Classroom Test of Scientific Reasoning (LCTSR). It is a two-tier, multiple-choice test with 24 items (Lawson, 2000). A two-tier multiple-choice item pair contains a question with some possible answer choices, followed by another question proving some possible reasons for the response to the previous question. All the answer choices were designed based on previous studies on student misconceptions with free response tests, interviews, and relevant literature (Treagust, 1995).

The scoring methods of LCTSR

In accordance with the LCTSR design, both questions in a two-tier, multiple-choice item pair must be correct in order for the students to receive credit (Lawson, 2000). According to Lawson's method, getting both questions wrong in an item pair indicates the lowest level of scientific reasoning, and getting both correct indicates the highest, but the level is not distinguished when getting only one of the questions correct (only the answer, or only the reasoning) (Lawson, 2000; Treagust, 1995).



However, two latent traits are embedded in the two-tier multiple-choice item pairs: answering the question is for knowing the result in the first tier, and choosing the reason is for explaining the reason in the second tier (Tsai & Chou, 2002). A number of studies have shown that explaining the reason represents a higher skill than answering the question correctly using the classical test theory (CTT) (Bayrak, 2013; Caleon & Subramaniam, 2009; 2010; Chang et al., 2007; Xiao et al., 2018). This implies that students may know the answer before they have developed the capability of explaining the reason for the answer. Therefore, there exist intermediate levels of understanding that correspond to different response patterns. For a two-tier, multiple-choice item pair, there are four different response patterns of correct and incorrect answers. Among these, "00" represents getting both answer and reasoning incorrect as the lowest level of student performance on the two-tier item pair, while "11" represents getting both answer and reasoning correct as the highest level. For the two intermediate patterns, "01" corresponds to incorrect answer with correct reasoning, which is often interpreted as guessing. On the other hand, "10", which corresponds to correct answer with incorrect reasoning may represent a higher level of understanding than guessing.

In general, three scoring methods are commonly employed for evaluating two-tier multiple-choice items in the previous studies: individual scoring method, pair scoring method, and partial credit scoring method. Individual scoring method treats two questions in a two-tier item pair as individuals and assigns score for each tier question independently (Chang et al., 2007; Chu et al., 2009). Pair scoring methods treats two questions in a two-tier item pair as a combined entity, and assigns credit only for answering both questions correctly, zero point for all other response patterns (Bayrak, 2013; Chandrasegaran et al., 2007; Lin, 2004). Following the assumption that students may know the answer to a question before they can fully articulate the reasoning based on their response, partial credit scoring method may assign 2 points for the pattern of "11", 1 point for the pattern of "10", and 0 point for patterns of "00" and "01" (Xiao et al., 2018), or assign 3 points for the pattern of "11", 2 point for the pattern of "10", 1 point for the pattern of "01" and 0 point for the pattern of "00" (Satriana et al., 2018; Xiao et al., 2018). Xiao et al. (2018) pointed out that individual scoring method rewards the intermediate levels, but it may also assign credit for guessing, while pair scoring method avoids guessing rewards, but ignores the possible intermediate learning stages by underlining the relationship of knowing the answer and explaining the reason. Rasch analysis was then used to explore different scoring methods on the data of LCTSR, and the results confirmed that for the partial credit scoring method, the pattern "10" represents a higher level of scientific reasoning than the pattern "01", which should be treated as guessing.

Research Questions

Among the previous studies on analyzing two-tier, multiple-choice items, some were limited to analyzing individual two-tier, multiple-choice item pair (Chang et al., 2007; Satriana et al., 2018; Xiao et al., 2018), while others treated different two-tier, multiple-choice item pairs as independent combinations, assigning scores for individual items pairs, and simply adding them together for a total score of all items (Lawson, 2000; Luo et al., 2020). However, the possible inclusion of several item pairs that probe the same aspect of scientific reasoning yet with varying difficulties is ignored. Neither examining the item pairs individually nor simply adding all their scores fully explores this dimension of item difficulty.

From the theory of cognitive development (Inhelder & Piaget, 1958; Piaget, 1971), the development of reasoning follows a process that evolves from simple to complex (Watson, 1975). Students are likely to understand simple phenomena, concepts, and laws of the nature first before some of the more difficult ones that are in the same domain. Similarly, students are likely to gain the ability to solve simple problems first before complicated problems in the same domain. Therefore, students' performance on item pairs of varying difficulty in the same domain can provide important information on student's scientific reasoning development.

Building on previous works, this research aimed to establish a method of pattern analysis for assessing two-tier item pairs in LCTSR. The method of pattern analysis was chosen for the research, as it is a common technique in data mining, where processes, algorithms, and mechanisms are investigated to retrieve potential knowledge from data collections (Norton, 1999). In this case, pattern analysis is an intuitive and straightforward method to investigate students' performance on item pairs of different difficult levels. Through the pattern analysis in student responses, the finer details of students' scientific reasoning development levels can be probed by examining a cross-section of all the student data with different grade levels. Specifically, this research aimed to answer three research questions:



1. Do intermediate response patterns represent different reasoning levels in terms of grade levels?
2. Do intermediate response patterns represent different reasoning levels in terms of students' overall development?
3. Are combined response patterns to item pairs of different difficulty good indicators of reasoning development levels?

Research Methodology

The Difficulty Characterization of Two-tier Item Pairs in LCTSR

When LCTSR was first designed, the items typically fell into three levels of difficulty: concrete reasoning, transitional reasoning, and formal reasoning (Lawson, 1978). For instance, for the test items designed to measure student reasoning on the subskill of Control-of-Variable (COV), two item pairs have been shown to have significant difference in difficulty (Lawson, 2000). These two item pairs are presented in Figure 1, and the individual items are labeled as P1 (pendulum answer), P2 (pendulum reasoning), F1 (flies in a tube answer), and F2 (flies in a tube reasoning).

The two responses for P1 and P2 are listed as the first pair, while those for F1 and F2 are listed as the second pair. Each pair has an answer item and a reason item. The truth value of the answer is the first digit in each pair, and that of the reasoning is the second digit. As a basic indicator of difficulty, the percent correct was calculated from the performance among students from middle school (grade 6-7), high school (grade 9-10), and college, using individual score method for these four individual items, and the results are shown in Table 1. As can be seen in Table 1, it is evident that students from all grade levels performed better on the first item pair (P1 and P2) than the second (F1 and F2), thus it can be assumed that the first item pair (P1 and P2) is easier than the second (F1 and F2). The present research focuses on students' performance on these four items (P1, P2, F1, and F2) when analyzing student reasoning on the subskill of Control-of-Variable (COV) of LCTSR.

Table 1

Percent Correct on the Two Item Pairs in LCTSR across Different Grade Levels

Item	Context	Percent correct (%)			Difficulty level	
		Grades 6-7	Grades 9-10	College		
1st item pair	P1	Pendulum – answer	34	66	79	Easy
	P2	Pendulum – reasoning	29	57	78	
2nd item pair	F1	Flies – answer	16	29	50	Difficult
	F2	Flies – reasoning	16	17	30	

Sample

Since students' scientific reasoning development levels were important to the present research, the participants were constituted as students in different grade levels. One part of the participants consisted of students in grade 4-12 in both the United States and China, and the other part included college first-year students from a large Midwestern university in the United States. The student responses from the United States and China were not distinguished in the analysis, since it has been shown that scientific reasoning levels of the two sets of population are effectively similar (Bao et al., 2009). Therefore, there were a total number of 10707 students enrolled into the present research with the number distribution of participating students shown in Table 2. Specifically, in the primary level, there were 336 students from grade four, 547 students from grade five and 588 students from grade six. In the junior school level, there were 868 students from grade seven, 606 students from grade eight, and 1489 students from grade nine. In the secondary school level, there were 1520 students from grade ten, 2083 students from grade eleven, and 847 students from grade twelve. In addition, 1823 college first-year students participated in the present research.

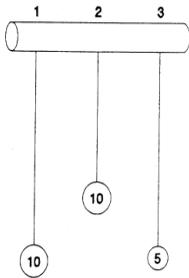


Table 2
The Number Distribution of Participating Students

Grade Level	4	5	6	7	8	9	10	11	12	College
N	336	547	588	868	606	1489	1520	2083	847	1823

Figure 1
Two-Tier, Multiple-Choice Item Pairs from LCTSR Used in this Research

P1
At the right are drawings of three strings hanging from a bar. The three strings have metal weights attached to their ends. String 1 and String 3 are the same length. String 2 is shorter. A 10 unit weight is attached to the end of String 1. A 10 unit weight is also attached to the end of String 2. A 5 unit weight is attached to the end of String 3. The strings (and attached weights) can be swung back and forth and the time it takes to make a swing can be timed.



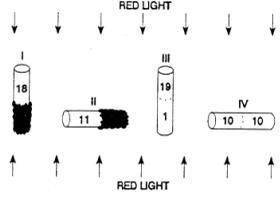
Suppose you want to find out whether the length of the string has an effect on the time it takes to swing back and forth. Which strings would you use to find out?

- only one string
- all three strings
- 2 and 3
- 1 and 3
- 1 and 2

P2
because

- you must use the longest strings.
- you must compare strings with both light and heavy weights.
- only the lengths differ.
- to make all possible comparisons.
- the weights differ.

F1
Twenty fruit flies are placed in each of four glass tubes. The tubes are sealed. Tubes I and II are partially covered with black paper; Tubes III and IV are not covered. The tubes are placed as shown. Then they are exposed to red light for five minutes. The number of flies in the uncovered part of each tube is shown in the drawing.



This experiment shows that flies respond to (respond means move to or away from):

- red light but not gravity
- gravity but not red light
- both red light and gravity
- neither red light nor gravity

F2
because

- most flies are in the upper end of Tube III but spread about evenly in Tube II.
- most flies did not go to the bottom of Tubes I and III.
- the flies need light to see and must fly against gravity.
- the majority of flies are in the upper ends and in the lighted ends of the tubes.
- some flies are in both ends of each tube.

Notes: P1:pendulum answer, easy; P2: pendulum reasoning, easy; F1: flies in a tube answer, difficult; F2: flies in a tube reasoning, difficult.

Data Collection

All students volunteered to take the LCTSR. For the administration of the test, students in the United States used the English version, while those in China used the Chinese version. To ensure the consistency between the two versions, the test was carefully translated and evaluated by a group of six faculty members who are proficient in both languages. All students were given sufficient time to finish the test. Students from lower grade levels took about 45 to 50 minutes for the test, while college students needed about 30 minutes. Prior to the formal implementation of the test, a sample test was conducted with both primary students and college freshmen to ensure that the time provided was appropriate.

Data Analysis

This research focused on students' performance on the four chosen items (P1, P2, F1, and F2) from LCTSR when analyzing student reasoning on COV. In each item pair, students choose an answer to the question in the first tier, and choose a reason of explanation in the second tier. Responses are coded using "0" for an incorrect answer and "1" for a correct answer. Thus, the code 00-00 means all responses are incorrect, while a code of 11-11 means all

responses are correct. A code of 11-10 means that the student provides correct responses to P1, P2, and F1, but an incorrect response to F2. Since each of the four items can be answered correctly or incorrectly, there are a total of 16 possible response patterns: 00-00, 00-01, 01-00, 01-01, 00-10, 01-10, 00-11, 01-11, 10-00, 10-01, 11-00, 11-01, 10-10, 10-11, 11-10, and 11-11.

As previous research studies have suggested that explaining the reason represents a higher level of reasoning than answering the question correctly (Bayrak, 2013; Caleon & Subramaniam, 2009; 2010; Chang et al., 2007; Xiao et al., 2018), students are likely to have the ability to provide a correct answer before they can clearly articulate correct reasoning. Therefore, students' intermediate levels of scientific reasoning may be expressed within the different response patterns on the four items chosen from LCTSR.

Another factor that can be examined to indicate students' level of scientific reasoning development is the item difficulty. Since P1 and P2 are considered to be easier than F1 and F2, students who can answer F1 and F2 correctly are considered to have a higher level of reasoning.

Based on the above assumptions, the 16 response patterns were analyzed and matched with different levels of reasoning. According to Piaget's cognitive development theory, children's cognitive development can be divided into four stages by age group (Inhelder & Piaget, 1958; Piaget, 1971): sensorimotor stage (ages 0-2), preoperational stage (ages 2-7), concrete operational stage (ages 7-12), and formal operational stage (ages 12 to adult). Therefore, a higher level of a particular subskill is more likely to be observed among students from higher grade levels and higher overall scientific reasoning skill levels. This assumption serves as the basis of the student performance analysis across different grade levels and overall scientific reasoning development levels.

Research Results

Intermediate Response Patterns Change with Students' Grade Levels

Student response patterns of the four COV items from grade 4 through college were collected and organized. Since the first item pair (P1 and P2) is easier than the second (F1 and F2), four intermediate response patterns are compared first, which include 01-00, 10-00, 11-01, and 11-10. The percentage of these four patterns at different grades are listed in Table 3.

Table 3
Student Performance on P1, P2, F1, and F2 from Grade 4 through College

Grade Levels	Patterns			
	01-00 (%)	10-00 (%)	11-01 (%)	11-10 (%)
4	9.5	6.5	0.6	0.3
5	6.4	9.0	0.9	1.1
6	6.0	10.7	2.4	1.5
7	5.1	8.2	2.0	3.1
8	2.8	17.0	1.5	10.1
9	3.8	12.0	3.2	9.9
10	4.6	9.5	2.9	12.8
11	3.5	8.5	4.6	14.5
12	1.8	4.4	4.1	23.1
College	0.4	1.6	4.4	21.5

The results indicated that as the grade level increases, the percentage of student responses of "01-00" decreases gradually. As for student response of "10-00", the percentage starts rising from grade 4, peaks in grade 8, and then falls gradually until college. From grade 4 to grade 8, the increase in the "10-00" pattern suggests an improving cognitive development with grade levels. The decrease after grade 8 is due to a larger portion of the students



moving to higher developmental level (i.e. answering both P1 and P2 correctly), which is an indication of a major learning shift, and a critical point in students' scientific reasoning development.

In the case where the students answer the first pair correctly, the percentage of student responses of "11-01" is relatively small and independent of grade levels. This is another evidence that "01" does not accurately reflect students' level of scientific reasoning development, and is likely a result of guessing. As for student response of "11-10", the percentage rises steadily as the grade level increases, from 0.3% in grade 4 to 21.5% in college. It indicates "10" as some level of reasoning development, which is higher than "00" or "01".

Putting all these together, this cross-sectional study of all students from grade 4 through college confirms that student responses of "11-10" and "10-00" represent higher levels of scientific reasoning development than "11-01" and "01-00", respectively.

Intermediate Response Patterns Change with Students' Overall Development

Student response patterns of the four COV items were analyzed based on their overall reasoning development levels. In order to eliminate any variation due to the grade level, three subsets of the whole sample were chosen based on the grade level, with at least two-grade-level difference between each group: grades 6-7, grades 9-10, and college.

Within each group, students were divided among three different overall reasoning development levels based on their total scores on the remaining 20 items of the LCTSR, excluding the four chosen COV items. These levels were defined based on Lawson's original research (Lawson, 1978), which included bottom 30% for concrete reasoning, medium 40% for transitional reasoning, and top 30% for formal reasoning. In this research, these levels translated approximately into score cutoffs (out of 20 points) at 0 to 11 for bottom 30%, 12 to 16 for the intermediate 40%, and 17 to 20 for the top 30%. The final percentages and the number of students in each development level for different populations are shown in Table 4.

Table 4

Responses to Lawson Test Items P1, P2, F1, and F2 from Grades 6-7, 9-10, and College

	Percentage of ranking (%)	Score	N	Patterns			
				01-00 (%)	10-00 (%)	11-01 (%)	11-10 (%)
Grades 6-7	Low 30	0-5	433	6.2	7.8	1.6	0.9
	Mid 42	6-9	616	5.8	8.6	1.6	1.1
	High 28	10-19	404	4.0	11.6	3.5	6.2
Grades 9-10	Low 33	0-8	991	7.8	11.3	2.4	3.1
	Mid 40	9-13	1211	2.6	11.8	3.5	10.2
	High 27	14-20	804	2.2	8.4	3.2	23.1
College	Low 29	0-11	523	1.1	3.4	4.2	9.2
	Mid 40	12-16	724	0.1	1.4	6.2	20.8
	High 31	17-20	573	0.2	0.2	2.3	33.6

Notes: Low (%) means the bottom part of students; Mid (%) means the intermediate part of students; and High (%) means the top part of students.

Similar to the previous analysis on grade levels, the intermediate response patterns on the two pairs of COV questions were examined. The results are also included in Table 4, which indicate that the percentage of student responses of "01-00" and "11-01" are relatively low, across all grade levels. Within each group, these percentages are not significant, and independent of the overall reasoning development level. Both of these two observations



suggest that the "01" type response does not represent a meaningful level of reasoning development, and is likely due to guessing.

On the other hand, the percentage of student responses of "10-00" and "11-10" are relatively high, across all grade levels. Moreover, within each group, as the level of overall reasoning increases, the percentage of student responses of "11-10" increases accordingly, contrary to that of "11-01". This relation is more pronounced in grades 9-10 and college. These observations also well explain that the "10" response indicates a meaningful level of scientific reasoning, while "01" response indicates guessing.

Putting all these together, the two pattern analysis on student responses of different grade levels and overall reasoning development suggest that a correct answer with incorrect reasoning indicates an intermediate level of reasoning development, whereas an incorrect answer with correct reasoning is likely a result of guessing. As a result, among the 16 answer patterns to the two-tier item pairs of LCTSR, the pattern "10-00" and "11-10" represent a higher level of scientific reasoning development than "01-00" and "11-01", respectively.

Combined Response Patterns to Item Pairs of Different Difficulty as Good Indicators of Reasoning Development Levels

As discussed previously, the response patterns of "01" and "10" represent different levels of development. Accordingly, students' development in COV can be divided into different levels based on their responses to the four chosen COV items. The pattern "00-00" represents the lowest level, whereas "11-11" represents the highest. The intermediate levels are ordered based on the rules shown below:

1. Students are able to provide a correct answer before they can provide the correct reasoning for the same item;
2. Students are able to answer the easy items correctly before they can answer the difficult ones correctly;
3. The "01" response (correct answer, incorrect reasoning) is likely a result of guessing.

Following these considerations, all 16 possible response patterns are grouped into six levels.

Level 1 (00-00) includes students giving all incorrect answers, and is the lowest level of reasoning development.

Level 2 (0x-xx) includes students giving an incorrect response to P1, which expands into 7 patterns: "00-01", "01-00", "01-01", "00-10", "01-10", "00-11", and "01-11". If a student cannot answer P1 correctly, according to the rules, it is likely that any other correct answers from other items are due to guessing. This level also includes the students giving incorrect responses to both P1 and P2 incorrect but correct one(s) to either or both of F1 and F2. Since the first item pair (P1 and P2) is easier than the second (F1 and F2), these responses are also considered as the results of guessing. Although level 2 is mostly comprised of guessing responses, it is still considered as a level higher than level 1, because a correctly guessed response may indicate some level of reasoning, however little or implicit. For instance, a student can eliminate some answer choices, and make an informed guess. On the other hand, the "00" response could indicate a misconception, which should be separated from guessing.

Level 3 (10-0x) includes students giving correct responses to P1, incorrect to P2, and incorrect to F1: namely, "10-00" and "10-01". The "10-01" response is included in this level, as the correct response of F2 is likely due to guessing. This level is considered higher than level 2 because the response to P1 is correct.

Level 4 (11-0x, 10-1x) includes students giving correct responses to both P1 and P2 or P1 and F1: "11-00", "11-01", "10-10" and "10-11". These responses are grouped together, as it is unclear which of these responses indicates a higher level of reasoning. The "10-11" response is included in this level because there is a possibility that students miss P2 while still fully understand F1 and F2. Also, it is unlikely that a student would guess both F1 and F2 correctly.

Level 5 (11-10) includes students giving correct responses to P1, P2, and F1, and incorrect responses to F2: "11-10". It means that students fully understand the easier item pair (P1 and P2), and are in midway to understand the more difficult item pair.

Level 6 (11-11) includes students giving correct responses to all items, which represents the highest level of development.



Table 5*Percentage of Response Patterns from Students in Different Developmental Stages of Scientific Reasoning*

Stages of reasoning development		Response Patterns						Row Sum (%)
		1	2	3	4	5	6	
Stages	Overall Score (%)	00-00 (%)	0x-xx (%)	10-0x (%)	11-0x, 10-1x (%)	11-10 (%)	11-11 (%)	
1	0.0	46.5	38.2	8.7	6.0	0.5	0.2	100.0
2	10.0	42.6	30.0	11.7	13.5	1.5	0.6	100.0
3	20.0	36.4	23.1	11.9	24.2	3.0	1.4	100.0
4	30.0	29.5	16.2	12.5	35.4	4.6	1.8	100.0
5	40.0	20.8	13.3	11.5	41.1	8.5	4.8	100.0
6	50.0	14.0	12.8	10.5	43.0	11.9	7.9	100.0
7	60.0	9.5	8.0	9.7	46.5	15.4	10.8	100.0
8	70.0	3.9	5.9	5.7	41.3	24.7	18.5	100.0
9	80.0	1.5	5.2	3.2	33.0	35.1	22.0	100.0
10	90.0	1.9	4.8	1.7	26.6	31.2	33.8	100.0
11	100.0	1.0	5.4	3.4	11.3	35.3	43.6	100.0

In order to explore the development of COV skills, all students were divided into 11 developmental stages according to the percentage of correct answers of the remaining 20 items in LCTSR. For each of the developmental stages, the percentages of students responding with the different patterns are calculated and given in Table 5. The changes of the patterns across different developmental stages are plotted in Figure 2.

The results show that as expected level 1 patterns start with high percentages in the low-end of reasoning development and decrease steadily as the stage of reasoning development increases. Similarly, the level 2 patterns also decrease as the reasoning development stage increases. This is also expected since the level 2 responses represent mostly the results of guessing, which should decrease as students' reasoning skills develop into higher stages.

The level 3 patterns are relatively steady for students from the low-end through the middle of the reasoning development stages. These students usually can respond to P1 (the first answer) correctly but have incorrect answers to P2 (the reasoning), confirming that answer often precedes reasoning. On the other hand, students in higher reasoning stages have less level 3 patterns, which suggests that students at higher reasoning stages have developed both "knowing" and "reasoning".

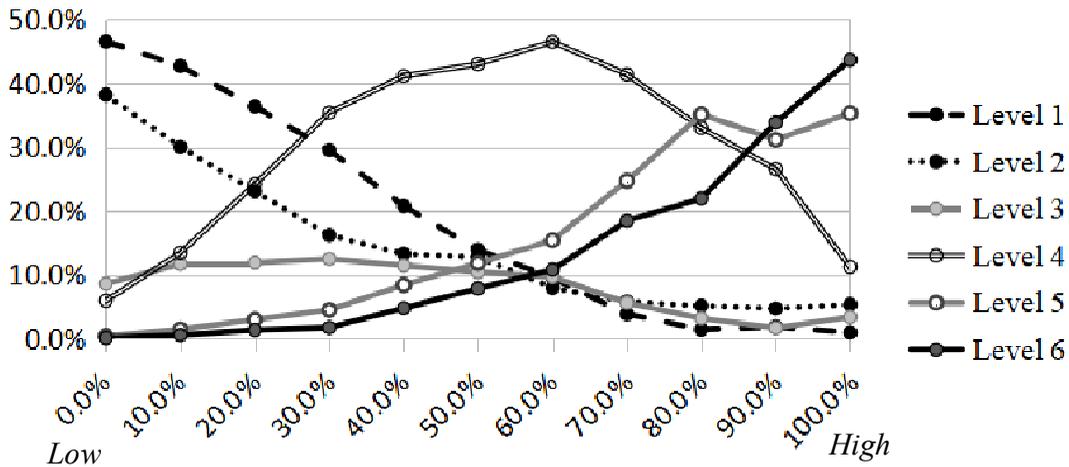
The level 4 patterns first increase then decrease with the developmental stage. The results show that more and more students are able to answer P1 and P2 correctly as their reasoning skills develop, which accounts for the initial increase. Meanwhile students at higher overall reasoning stages begin to answer F1 and F2 correctly, which accounts for the decrease.

The level 5 patterns also increase with the reasoning development stage and somewhat plateaued at the highest stages. Meanwhile the level 6 patterns increase steadily with the developmental stage. The increases of both patterns are more dramatic after stage 7, which suggest that a thorough understanding of COV skills are achieved when students develop into the formal reason stage (top 30%).



Figure 2

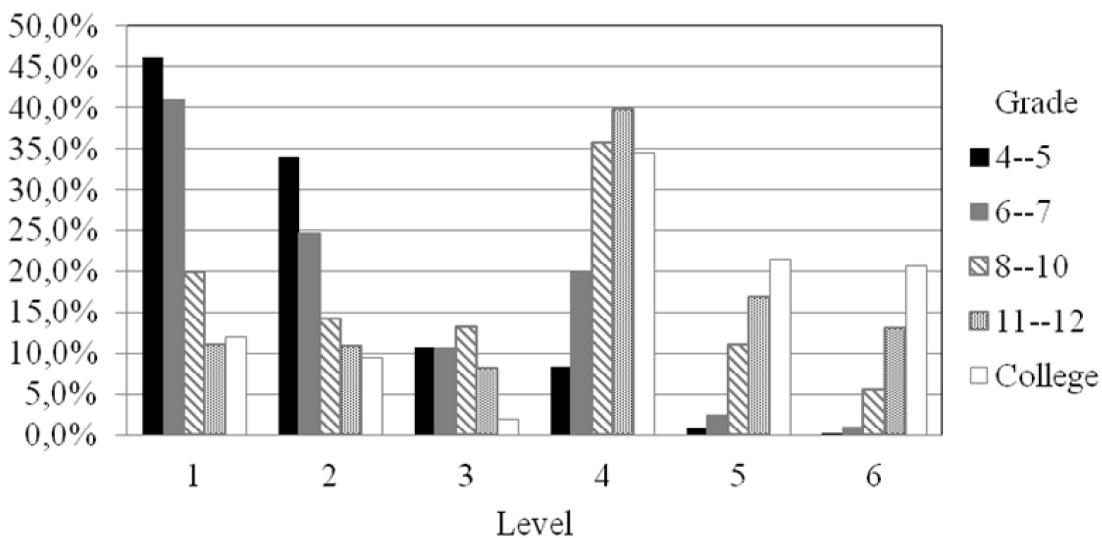
Percentage of Students From Low-End to High-End Overall Scientific Reasoning Abilities at the Six Performance Levels On P1, P2, F1, And F2 Of LCTSR



The results in Figure 2 illustrate the trends of how the different response patterns change with students' stages of overall reasoning development. Among the six levels of response patterns, the most interesting one is level 4, which shows a clear concave down shape with a peak at around 60% of the overall reasoning score. The remaining five levels are mostly monotonically increasing or decreasing. This suggests that level 4 patterns can be an important transitional stage of learning that students may first develop into and then evolve beyond. Therefore, assessment of the level 4 patterns may provide useful indicators of substantive cognitive transitions during the progression of learning.

Figure 3

Distribution of the Six Levels of Response Patterns across Different Grades



The same data can also be analyzed in terms of grade level. Student responses are combined into 5 groups based on their grade levels: grades 4-5, 6-7, 8-10, 11-12, and college. The percentages of different levels of response patterns for different grade groups are plotted in Figure 3, which shows that the higher level of response patterns are more popular among advanced students. These results are consistent with those presented in



Table 5 and Figure 2. In particular, the level 4 patterns show a more pronounced peak among middle school to college students (grades 8-12 and college), which indicates that many of students in these grades are in the transitional stage of developing COV skills. This result also corroborates with previous analysis showing that the level 4 patterns can be indicators of significant learning transitions.

Discussion

In this research, a method of combined pattern analysis was proposed for evaluating two-tier item pairs on students' COV skills in LCTSR. To be able to measure students' COV skills in finer detail, combined patterns of two item pairs responses were analyzed to show how they correspond to students both in different grade levels and in various stages of overall reasoning development. A few key results were uncovered in this analysis. The essential goal of the research was to identify if getting just the answer or just the reasoning correct indicated different skill levels, and which represented a higher skill level. On the one hand, the research of how intermediate response patterns represent different reasoning levels in terms of grade levels supports that the "10" response indicates a higher level of scientific reasoning development than the "01" response by cross-sectional analysis of data from grade 4 through college. The result that the percentage of student responses of "01-00" decreases gradually as the grade level increases agrees with Piaget's cognitive development theory, inferring that a higher level on a specific subskill development is more likely to be observed among students from higher grade levels (Inhelder & Piaget, 1958; Piaget, 1971). On the other hand, detailed analysis of data from students in different overall reasoning development levels also suggests that a correct answer with incorrect reasoning indicates an intermediate level of reasoning development, whereas an incorrect answer with correct reasoning does not represent a meaningful level of reasoning development and is likely a result of guessing. These results support previous research studies that students are able to provide the correct answer before they can provide the correct reasoning (Bayrak, 2013; Caleon & Subramaniam, 2009; 2010; Chang et al., 2007; Xiao et al., 2018). This is a result that is often overlooked by traditional two-tier scoring methods, in which the "01" and "10" responses are not explicitly distinguished. The traditional scoring of the Lawson Test allows for only two levels of performance, both the answer and reasoning need to be correct or no credit is given. It has been proved in the present research that the traditional scoring method does not accurately reflect the possible levels of student understanding. Students who get just the answer correct should be at a higher level of understanding than those who get just the reasoning correct or who get both incorrect. It is believed that students' skill levels should be identified at a finer grain size. This leads to a step-function in the scoring of a particular individual question or two-tier question.

Based on the previous work, sixteen combined response patterns to the two-tier item pairs of LCTSR were ordered and six performance levels were established based on some proven rules. Aside from Piaget's cognitive development theory that students are able to answer the easy items correctly before they can answer the difficult ones correctly (Inhelder & Piaget, 1958; Piaget, 1971), the rules also included that students are able to provide a correct answer before they can provide the correct reasoning, and an incorrect answer with correct reasoning is likely a result of guessing. From the results, student performance resembles a developmental process from low to high skill levels, with level 1 in the low-end and level 6 in the high-end of reasoning development. With the method of pattern analysis, a relation was obtained between students' COV skill, grade level, and overall reasoning development. It can be found how the trends of the different response patterns change with students' grade levels and their stages of overall reasoning development. Particularly, level 4 patterns can be an important transitional stage of learning that students may first develop into and then evolve beyond. The level 4 patterns with a pronounced peak indicate that many of students among middle school to college students are in the transitional stage of developing COV skills. It suggests that the level 4 patterns provide useful indicators of significant learning transitions. As the data have shown, reasoning skills develop slowly, and there is an intermediate level that traditional scoring does not recognize. The present result is highly valuable, indicating that pattern analysis could reflect students' ability more accurately than the traditional scoring of two-tier questions, contribute to better data analysis, and provide with a simple way to track learning progress.

The research strives to strike a balance between the complexity of the data analysis and the simplicity of grading a multiple-choice assessment instrument. Through the above efforts, students can perceive that they still retain some valuable thinking even though they may not solve problems correctly in all aspects. This



would provide motivation and a sense of achievement for students in the process of developing cognitive progression. The results also enable teachers and students to intuitively realize the unique meaning between “knowing” and “reasoning”. As in Moraes et al. (2020)’ proposal of using Problem Based Learning (PBL) method to cultivate students’ scientific literacy and science concepts, the teachers first consider how students arrive at the conclusions to the questions and then how the systemization of learning can be achieved. Even students’ scientific reasoning were demonstrated to be improved at high, moderate or low levels for different dimensions (Erlina et al., 2018), pattern analysis could be served as a technique to deeply excavate student performance of the intermediate level difference for each dimension.

Conclusions

The present research first conducted a cross-sectional analysis of data from grade 4 through college in both the United States and China to explore how intermediate response patterns represent different reasoning levels in terms of both students’ grade levels and overall reasoning development levels. The results showed that a correct answer with incorrect reasoning indicates an intermediate level of reasoning development, whereas an incorrect answer with correct reasoning is likely a result of guessing. Based on the combined response patterns, six performance levels for the two-tier item pairs were established to reflect students’ developmental process from low to high skill levels. In particular, the level 4 patterns provide useful indicators of substantial cognitive transitions during the learning progression.

The pattern analysis method provides a new way to exploit the information embedded within the combined response patterns of two two-tier item pairs of different difficulties. The analysis outcomes are consistent with Piaget’s cognitive development theory that a higher level of a particular reasoning skill development is more likely to be observed among students from higher grade levels and higher overall scientific reasoning development levels. The pattern analysis is an easily accessible method and provides a straightforward interpretation of the data. Therefore, this approach can provide a supplemental means to standard statistical tools, which can reveal useful information to enrich the interpretation of assessment data regarding students’ learning and development. Future work with this pattern analysis method can specialize in additional item pairs in LCTSR, as well as other scientific reasoning instruments for their potential power in defining the scientific reasoning learning progression.

Acknowledgements

The research is supported in part by 2020 Guangzhou Philosophy and Social Science Planning Fund of P.R. China under the Grant No. 2020GZQN20, and by the National Science Foundation Award DUE-1712238. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

References

- Adey, P., & Shayer, M. (1990). Accelerating the development of formal thinking in middle and high school students. *Journal of Research in Science Teaching*, 27(3), 267-285. <https://doi.org/10.1002/tea.3660270309>
- Bao, L., Cai, T., Koenig, K., Fang, K., Han, J., Wang, J., Liu, Q., Ding, L., Cui, L., Luo, Y., Wang, Y., Li, L., & Wu, N. (2009). Learning and scientific reasoning. *Science*, 323(5914), 586-587. <https://doi.org/10.1126/science.1167740>
- Bayrak, B. K. (2013). Using two-tier test to identify primary students’ conceptual understanding and alternative conceptions in acid base. *Mevlana International Journal of Education*, 3(2), 19-26.
- Caleon, I. S., & Subramaniam, R. (2009). 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. *Research in Science Education*, 40(3), 313-337. <https://doi.org/10.1007/s11165-009-9122-4>
- Caleon, I. S., & Subramaniam, R. (2010). Development and application of a three-tier diagnostic test to assess secondary students’ understanding of waves. *International Journal of Science Education*, 32(7), 939-961. <https://doi.org/10.1080/09500690902890130>
- Chandrasegaran, A. L., Treagust, D. F., & Mocerino, M. (2007). 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. *Chemistry Education Research & Practice*, 8(3), 293-307. <https://doi.org/10.1039/B7RP90006F>
- Chang, H.-P., Chen, J.-Y., Guo, C.-J., Chen, C.-C., Chang, C.-Y., & Lin, S.-H., et al. (2007). Investigating primary and secondary students' learning of physics concepts in Taiwan. *International Journal of Science Education*, 29(4), 465-482. <https://doi.org/10.1080/09500690601073210>
- Chu, H.-E., Treagust, D. F., & Chandrasegaran, A.L. (2009). A stratified study of students' understanding of basic optics concepts in different contexts using two-tier multiple-choice items. *Research in Science & Technological Education*, 27(3), 253-265. <https://doi.org/10.1080/02635140903162553>
- Coletta, V. P., & Phillips, J. A. (2005). Interpreting FCI scores: Normalized gain, reinstruction scores, and scientific reasoning ability. *American Journal of Physics*, 73(12), 1172-1179. <https://doi.org/10.1119/1.2117109>
- Dunbar, K., & Klahr, D. (2012). Scientific thinking and reasoning. In K. Holyoak & R. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 701-718). Oxford University Press.
- Erlina, N., Susantini, E., Wasis, W., & Pandiangan, P. (2018). The Effectiveness of evidence-based reasoning in inquiry-based physics teaching to increase students' scientific reasoning. *Journal of Baltic Science Education*, 17(6), 972-985. <https://doi.org/10.33225/jbse/18.17.972>
- Goldschmid, M. L. (1967). Different types of conservation and non-conservation and their relation to age, sex, IQ, MA, and vocabulary. *Child Development*, 38(4), 1229-1246. <https://doi.org/10.2307/1127120>
- Inhelder, B., & Piaget, J. (1958). *The grow of logical thinking*. Basic Books.
- iSTAR Assessment. (2010). *iSTAR Assessment: Inquiry for scientific thinking and reasoning*. <http://www.istarassessment.org/>
- Joep, V. D. G., Eva, V. D. S., Gijssels, M., & Segers, E. (2019). A combined approach to strengthen children's scientific thinking: Direct instruction on scientific reasoning and training of teacher's verbal support. *International Journal of Science Education*, 41(9), 1119-1138. <https://doi.org/10.1080/09500693.2019.1594442>
- Johnson, M. A., & Lawson, A. E. (1998). What are the relative effects of reasoning ability and prior knowledge on biology achievement in expository and inquiry classes? *Journal of Research in Science Teaching*, 35(1), 89-103. [https://doi.org/10.1002/\(SICI\)1098-2736\(199801\)35:1<89::AID-JRST1098>3.0.CO;2-J](https://doi.org/10.1002/(SICI)1098-2736(199801)35:1<89::AID-JRST1098>3.0.CO;2-J)
- Lawson, A. E. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching*, 15(1), 11-24.
- Lawson, A. E., (2000). Classroom test of scientific reasoning: Multiple choice version. Based on a. E. Lawson, "development and validation of the classroom test of formal reasoning". *Journal of Research in Science Teaching*, 5(1), 11-24.
- Lawson, A. E., & Blake, A. J. D. (1976). The factor structure of some Piagetian tasks. *Journal of Research in Science Teaching*, 13(5), 461-466.
- Lawson, A. E., Nordland, F. H., & Kahle, J. B. (1975). Levels of intellectual development and reading ability in disadvantaged students and the teaching of science. *Science Education*, 59(1), 113-126.
- Lin, S. W. (2004). Development and application of a two-tier diagnostic test for high school students' understanding of flowering plant growth and development. *International Journal of Science and Mathematics Education*, 2(2), 175-199. <https://doi.org/10.1007/s10763-004-6484-y>
- Luo, M., Wang, Z., Sun, D., Wan, Z., & Zhu, L. (2020). Evaluating scientific reasoning ability: The design and validation of an assessment with a focus on reasoning and the use of evidence. *Journal of Baltic Science Education*, 19(2), 261-275. <https://doi.org/10.33225/jbse/20.19.261>
- Moraes, J. V., Castellar, S. M. V., Castellar, S. V., & Castellar, S. V. (2010). Scientific literacy, problem-based learning and citizenship: A suggestion for geography studies teaching. *Problems of Education in the 21st Century*, 19, 119-127. <http://www.scientiasocialis.lt/pec/node/364>
- Norton, M. J. (1999). Knowledge discovery in databases. *Library Trends*, 48(1), 9-21.
- Piaget, J. (1971). The theory of stages in cognitive development. In D. R. Green, M. P. Ford, & G. B. Flamer (Eds.), *Measurement and Piaget*. McGraw-Hill.
- Pratt, C., & Hacker, R. G. (1984). Is Lawson's Classroom Test of Formal Reasoning Valid? *Educational and Psychological Measurement*, 44(2), 441-448. <https://doi.org/10.1177/0013164484442025>
- Satriana, T., Yamtinah, S., Ashadi, & Indriyanti, N. Y. (2018). Student's profile of misconception in chemical equilibrium. *Journal of Physics: Conference Series*, 1097, 012066. <https://doi.org/10.1088/1742-6596/1097/1/012066>
- Stefanich, G. P., Unruh, R. D., Perry, B., & Phillips, G. (1983). Convergent validity of group tests of cognitive development. *Journal of Research in Science Teaching*, 20(6), 557-563.
- Treagust, D. F. (1995). Diagnostic assessment of students' science knowledge. In S. M. Glynn & R. Duit (Eds.), *Learning science in the schools: Research reforming practice* (pp. 327-346). Lawrence Erlbaum Associates.
- Tsai, C. C., & Chou, C. (2002). Diagnosing students' alternative conceptions in science. *Journal of Computer Assisted Learning*, 18(2), 157-165.
- Watson, M. S. (1975). A developmental study of empathy: Egocentrism to sociocentrism or simple to complex reasoning. *Cognitive Development*, 15, 1-14. <https://files.eric.ed.gov/fulltext/ED114179.pdf>
- Xiao, Y., Han, J., Koenig, K., Xiong, J., & Bao, L. (2018). Multilevel Rasch modeling of two-tier multiple-choice test: A case study using Lawson's classroom test of scientific reasoning. *Physical Review Physics Education Research*, 14, 020104. <https://doi.org/10.1103/PhysRevPhysEducRes.14.020104>



- Zhou, S., Han, J., Koenig, K., Raplinger, A., Pi, Y., Li, D., Fu, Z., & Bao, L. (2016). Assessment of scientific reasoning: The effects of task context, data, and design on student reasoning in control of variables. *Thinking Skills and Creativity*, 19, 175-187. <https://doi.org/10.1016/j.tsc.2015.11.004>
- Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, 27(2), 172-223. <https://doi.org/10.1016/j.dr.2006.12.001>

Received: July 06, 2020

Accepted: January 12, 2021

Cite as: Zhou, S.-N., Liu, Q.-Y., Koenig, K., Li, Q.-Y., Xiao, Y., & Bao, L. (2021). Analysis of two-tier question scoring methods: A case study on the Lawson's classroom test of scientific reasoning. *Journal of Baltic Science Education*, 20(1), 146-159. <https://doi.org/10.33225/jbse/21.20.146>

-
- Shao-Na Zhou** PhD, Associate Professor, School of Physics and Telecommunication Engineering, South China Normal University, Guangzhou 510006, China.
E-mail: zhou.shaona@m.scnu.edu.cn
ORCID: <https://orcid.org/0000-0003-1455-5122>
- Qiao-Yi Liu** PhD, Department of Physics, The Ohio State University, Columbus, OH 43210, USA.
mail: liu.6530@osu.edu
ORCID: <https://orcid.org/0000-0002-1613-7891>
- Kathleen Koenig** Associate Professor, Department of Physics, University of Cincinnati, Cincinnati, OH 45221, USA.
E-mail: kathy.koenig@uc.edu
- Qiu-ye Li** MSc, School of Physics and Telecommunication Engineering, South China Normal University, Guangzhou 510006, China.
E-mail: 2019021855@m.scnu.edu.cn
- Yang Xiao** PhD, School of Physics and Telecommunication Engineering, South China Normal University, Guangzhou 510006, China.
mail: 20092305002@m.scnu.edu.cn
ORCID: <https://orcid.org/0000-0002-2571-7759>
- Lei Bao**
(Corresponding author) PhD, Professor, Department of Physics, The Ohio State University, Columbus, OH 43210, USA.
E-mail: bao.15@osu.edu
ORCID: <https://orcid.org/0000-0003-3348-4198>
-

