International Journal of Assessment Tools in Education: Vol. 3, Issue 2 (2016) pp. 101-122

http://www.ijate.net
e-ISSN: 2148-7456 © IJATE
INTERNATIONAL JOURNAL of ASSESSMENT TOOLS in EDUCATON

"Original Research Article"

The Influence of Item Formats when Locating a Student on a Learning Progression in Science

Jing Chen[1,a], Amelia Wenk Gotwals[b], Charles W. Anderson[b], Mark D. Reckase[b]

[a]Educational Testing Service, Princeton, NJ08541, United States
[b]School of Education, Michigan State University, United States

Abstract

Learning progressions are used to describe how students' understanding of a topic progresses over time. This study evaluates the effectiveness of different item formats for placing students into levels along a learning progression for carbon cycling. The item formats investigated were Constructed Response (CR) items and two types of two-tier items: (1) Ordered Multiple-Choice (OMC) followed by CR items and (2) Multiple True or False (MTF) followed by CR items. Our results suggest that estimates of students' learning progression level based on OMC and MTF responses are moderately predictive of their level based on CR responses. With few exceptions, CR items were effective for differentiating students among learning progression levels. Based on the results, we discuss how to design and best use items in each format to more accurately measure students' level along learning progressions in science.

Article Info

Received
01January 2016

Revised:
16 March 2016

Accepted
20 March 2016

Keywords:
Item Format, Item Response Theory (IRT), Learning Progression, Science Assessment

1. INTRODUCTION

The goal of this study is to examine how to use different formats of items to classify students into learning progression levels, minimizing the measurement error of the assessments. Learning progressions are used to describe how students' understanding of a topic progresses over time (Corcoran, Mosher & Rogat, 2009). The National Research Council (NRC) recommends using learning progressions to inform the design and scoring of science assessments (NRC, 2006, 2007, 2014). Research suggests that assessment instruments that are developed in coordination with learning progressions can provide more information about a larger range of students than typical assessments (Songer, Kelcey & Gotwals, 2009; Songer & Gotwals, 2012) and offer more discriminatory power than traditional items (Liu, Lee, Hofstedder, & Linn, 2008).

However, learning progression-based assessments pose development challenges. One particular challenge is that it is difficult to write items that provide opportunities for students to

[1]Corresponding Author Email: jingchen@msu.edu

The study was conducted while the first author was at Michigan State University.

2148-7456 /© 2016 DOI: 10.21449/ijate.245196

respond at multiple levels of a learning progression (Anderson, Alonzo, Smith, & Wilson, 2007). Writing items may be especially difficult when the learning progression covers a broad range and there are large shifts in student understanding between levels of the learning progression. One way to combat this challenge is to use different types of assessment items. Different types of items may allow opportunities for students at multiple levels along a learning progression to respond in ways that indicate their level of understanding. Having a system that considers the construct, item types, and the appropriate measurement model from the beginning of the assessment design process can help to ensure that students at multiple levels along the learning progression have appropriate opportunities to respond (e.g., see Adams, Wilson, & Wang, 1997; van der Linden & Hambleton, 1996; Embretson, 1996; Songer, et al., 2009).

In this study, we examine the effectiveness of three item formats in classifying students' learning progression levels. Specifically, we examine Constructed Response (CR) items and two types of two-tier items; (1) Ordered Multiple-Choice (OMC; Briggs, Alonzo, Schwab, & Wilson, 2006) with CR items (OMC + CR) and (2) Multiple True or False (MTF) with CR items (MTF + CR). We use Item Response Theory (IRT) based analysis and descriptive statistics to evaluate the quality of these items and suggest ways to improve items in each format to classify students more precisely into learning progression levels.

1.1. Item Formats in This Study

Different item formats have advantages and disadvantages (Schuwirth & van der Vleuten, 2004). An important question is how to design a test composed of items in different formats to utilize the advantages of each format to measure a learning progression (National Assessment Governing Board [NAGB], 2010). Our CR items require students to respond to an open-ended item where they are not prompted with any distractors or ideas (see an example of a CR item below). These CR items are short-answer items, each of which requires about five minutes for students to answer. Each CR question is scored according to a scoring rubric that gives varying degrees of credit according to the learning progression achievement levels (see the example below).

An Example CR item (TREDEAB):

A tree falls in the forest. After many years, the tree will appear as a long, soft lump on the forest floor. The lump on the forest floor weighs less than the original tree. What happened to it? Where would you find the matter that used to be in the tree?

The major advantage of CR items is that they are more appropriate for measuring students' abilities to organize, integrate, and synthesize their knowledge and skills to solve novel problems. CR items can be used to demonstrate students' original thoughts and they allow students to show the process of their reasoning (Martinez, 1999; McNeill & Krajcik, 2007; Champagne, Kouba & Gentiluomo, 2008). Research suggests that CR items provide greater measurement precision at the high and low ends of the ability distribution (Ercikan, Schwarz, Julian, Burket, Weber, & Link, 1998; Lee, Liu, & Linn, 2011; Wilson & Wang, 1995). Ercikan, et al. (1998) discovered that when CR items and multiple choice items were combined to produce scores on a single scale, the overall measurement accuracy improved because the CR items could tap very-low and very-high ability groups. Wilson and Wang (1995) reported that "performance-based items provided more information than multiple-choice items and also provided greater precision for higher levels of the latent variable" (p. 51).

However, CR items have their own problems such as the difficulty in administrating, scoring, inconsistencies among raters, and not always showing students' thinking (Wainer & Thissen, 1993). In addition, the scores from CR items are often not very reliable because fewer CR items can be administered within a fixed amount of time (Thissen & Wainer, 1993) comparing to selected-response items. Thus, our assessment includes other item formats in addition to the CR format to measure students' learning progression of science.

The OMC + CR items are two-tier items that have two parts. The first part is an OMC question that requires students to choose an option that is linked to a particular learning progression level of students' understanding of the target concept. Students may receive partial credit if they select a response that represents lower or middle level understanding. The second part is a CR question that asks students to explain the choice they made in the OMC part (see an example OMC+CR item below). It may provide richer information about a student's understanding from reading the full response that explains his/her choice.

An Example OMC+CR Item (ACORN):

A small acorn grows into a large oak tree. Where does most of the weight of the oak tree come from? (Circle the best explanation from the list below).

a. From the natural growth of the tree. (level 1)
b. From carbon dioxide in the air and water in the soil. (level 3)
c. From nutrients that the tree absorbs through its roots. (level 2)
d. From sunlight that the tree uses for food. (level 1)

Explain why you think that the answer you chose is the best answer.

OMC items generally require relatively shorter administration time and less scoring effort than CR items. Advocates suggest that OMC items provide more diagnostic information than traditional multiple-choice items without sacrificing the efficiency advantages, and that results from OMC items can be used to communicate effectively about student understanding. One study (Briggs et al., 2006) suggested that test scores based on OMC items compared favorably with scores based on traditional multiple-choice items in terms of their reliability. Alonzo & Steedle (2008) found that compared to CR items, OMC items "appear to provide more precise diagnoses of students' learning progression levels and to be more valid, eliciting students' conceptions more similarly to cognitive interviews compared to open-ended items" (Alonzo & Steedle, 2008, p.1). However, the limitations of selected-response items are also well recognized such as the possibility for guessing, not eliciting students' original thoughts (Flowers, Bolton, & Brindle, 2008; Taleto-Miller, Han, & Guo, 2011), and might not be able to measure high order thinking (Delandshere & Petrosky, 1998; Kennedy, 1999; Lane, 2004).

The MTF + CR items are also two-tier items. The MTF part has a set of true or false questions and the CR part asks students to explain the choices they made in the MTF part (see an example MTF+CR item below). Compared to traditional multiple-choice items in which the test taker is asked to choose the best answer, MTF items allow for multiple answers and provide test takers more freedom to choose.

An Example MTF+CR Item (ENERPLNT):

Which of the following is (are) energy source(s) for plants? Circle yes or no for each of the following.

a. Water	YES NO
b. Light	YES NO
c. Air	YES NO
d. Nutrients in soil	YES NO
e. Plants make their own energy	YES NO

Please explain ALL your answers, including why the things you circled "No" for are NOT sources of energy for plants.

Frisbie (1992) gave a comprehensive review of the literature of the MTF format and synthesized the following merits of MTF items: (a) They are a highly efficient format for gathering achievement data, (b) they tend to yield more reliable scores than multiple-choice and other objective formats, (c) they measure the same skills and abilities as content-parallel multiple-choice items, (d) they are a bit harder than multiple-choice items for examinees, and (e) they are perceived by examinees as harder but more efficient than multiple-choice items (p. 25).

There are also shortcomings associated with the MTF format. Answering True or False items often involves guessing (Downing & Yudkowsky, 2009). Grosse and Wright (1985) found that the examinees' response style (guess "T" more often or guess "F" more often) would determine whether the scores from the questions keyed true or the scores from the questions keyed false were more reliable (e.g. if an examinee always guess "True" when he/she does not know the answer, then all of the error due to guessing is in the scores from the questions keyed true). Dunham (2007) found that students' responses to the MTF item were influenced by an "optimal number correct" response set. For example, examinees tended to endorse three or four of the six MTF options more frequently than would be expected by chance alone. These results suggest that MTF item can be used as an alternative to MC items, but when designing and analyzing MTF items, attention needs to be paid to the reliability of the items, the guessing involved in the responses, and the response style factor.

Although the advantages and disadvantages of each item format have been extensively discussed in the literature, little research has examined the effectiveness of each item type in differentiating students among learning progression levels. This study attempted to fill the gap by evaluating the effectiveness of each item format in distinguishing students among learning progression levels. The analysis of the two-tier items consisting of OMC and CR parts or consisting of MTF and CR parts helps to explore new forms of assessment to be used in classroom and large-scale contexts. Large-scale tests usually have constraints of money and time and tend to predominantly rely on selected-response items. To design large-sale tests that measure students' learning progression of a science topic, we need to design selected-response items such as OMC and MTF items that can diagnose students' achievement levels as well as CR or other performance based tasks.

1.2. Learning Progression Based Science Assessment

1.2.1. Assessing learning progressions

Learning progressions have become popular within the science education community because of their potential to build a bridge between research on how people learn, policy, and

methods for teaching and assessing science (Salinas, 2009; Corcoran et al., 2009). By tracing students' progress over time, researchers receive richer information about how students' understanding progresses and the pathways that they take in developing more sophisticated understandings. Learning progressions on different science topics have been proposed and verified (e.g., see Smith, Wiser, Anderson, & Krajcik, 2006; Merritt, Krajcik, & Shwartz, 2008; Alonzo & Steedle, 2008; Catley, Lehrer, & Reiser, 2004; Chen & Anderson, 2015).

The grain size of these learning progressions ranges from very small (e.g., over a single unit on force and motion; Alonzo & Steedle, 2008) to very large (e.g., the carbon cycling learning progression used in this study that characterizes student learning from elementary school through high school or beyond). Developing assessment items for smaller grain size learning progressions requires the ability to distinguish small changes in student understanding (e.g., see Rivet & Kastens, 2012). Thus, assessments need to be carefully designed to reduce measurement error and capture small differences. A test with a combination of selected-response items and CR items is likely to serve this purpose. The selected-response items can establish the reliability of the test since a large number of items can be administrated within a fixed amount of time and the CR items can diagnose deeper levels of understanding to measure students precisely.

Developing assessment items for larger grain size learning progressions is also challenging because it requires items that can elicit responses from students at all levels of the learning progression. Writing items that provide opportunities for students to respond at multiple levels of a learning progression is difficult (Anderson et al., 2007). For example, "more sophisticated students may not exhibit higher levels of understanding if the questions do not prompt a sophisticated response" (Alonzo & Gotwals, 2012, p. 248), but the types of questions that prompt a sophisticated response may include language or other features that are confusing to younger or less sophisticated students. A test composed of items in different formats may elicit responses from students at different levels and reduce measurement error for students over a wide range of abilities (e.g. Ercikan, et al., 1998). Thus, we investigate how to design and use items in different formats to accurately classify students into learning progression levels.

1.2.2. Items in multiple formats based on a learning progression of carbon cycling

The items we evaluated in this study were developed based on a learning progression about carbon cycling for which validity evidence is reported in prior research (Mohan, Chen & Anderson, 2009; Jin & Anderson, 2012). Data collected from hundreds of written assessments and dozens of clinical interviews suggest that students typically follow this learning progression when they progress from elementary to high school (Doherty, Draney, Shin, Kim, & Anderson, 2015). This learning progression includes four achievement levels presented in Table 1. The levels move from the lower anchor where students explain phenomena using "force-dynamic" reasoning to an upper anchor in which students are able to trace matter and energy systematically in their explanations. Our CR items are designed to elicit responses from students across different developmental levels. Students' responses to the CR items are classified into these four achievement levels.

Each option in an OMC question is linked to one of these four developmental levels of student's understanding, facilitating diagnostic interpretation of students' responses. For example, in the OMC question showed previously, option "a" and "d" represent understanding at level 1 in which students' understanding is confined to the macroscopic scale.

Table 1. Carbon Cycling Learning Progression

Level	Description
4	Students can use atomic-molecular models to trace matter/energy systematically through multiple processes connecting multiple scales. They use constrained principles (conservation of atoms and mass, energy conservation and degradation), codified representations (e.g. chemical equations, flow diagrams) to explain chemical changes.
3	Students can reason about macroscopic or large-scale phenomena but because of limited understanding at the atomic-molecular scale, they cannot trace matter and energy separately and consistently through those phenomena.
2	Students continue to attribute events to the purposes and natural tendencies of actors, but they also recognize that macroscopic changes result from "internal" or "barely visible" parts and mechanisms that involving changes of materials and energy in general.
1	Students describe the world in terms of objects and events rather than chemically-connected processes. Their understandings are confined to the macroscopic scale without recognizing the underlying chemical changes or energy transformations of events.

They think the weight gain of the oak tree comes from natural growth of the tree or from sunlight, which is recognizable to them. Option "c" represents understanding at level 2, in which students begin to recognize the weight of tree comes from invisible things such as nutrients. Option "b" is linked to the highest-level understanding among all the options. Students who select this option understand that carbon dioxide and water contribute most of the weight gain of the oak tree. It is worth noting that the options of OMC questions may not cover all the learning progression levels and sometimes there are multiple options that represent understanding at the same level (e.g. option 'a' and 'd').

The MTF questions provide students a list of things and ask them to judge whether each thing in the given list is the matter or energy source for events such as tree growth or human growth. The pattern of students' True or False choices represent understanding at different developmental levels. The example MTF+CR introduced previously is about energy and plants (item name is abbreviated as "ENERPLNT"). In this item, "light" is the only correct choice. Students who have a sophisticated understanding of energy transformation should be able to select True for light and False for all the other options. The options such as "water," "air," "nutrients." and "plants make their own energy" are lower level distractors. Students whose understanding is at intermediate or low levels are likely to choose these distractors that represent common misconceptions about the energy source of tree growth. The design of MTF options is grounded by learning progression levels.

Two research questions guided this study:

1. How effective are the OMC, MTF and CR item formats in classifying students' into learning progression levels?
2. What is the optimal use of OMC, MTF and CR items in designing a test to precisely measure students' learning progression levels?

2. METHODOLOGY

2.1. Participants

The written assessments were administered in twelve science teachers' classes from ten rural and suburban elementary, middle, and high schools in US Michigan, with four teachers at each level during 2009 to 2010 (see Table 2). The majority of the students in these schools are white. In total, 1,500 test papers were collected from their classes, including 316 at elementary, 727 at middle school, and 457 at high school. Almost all of the students in these teachers' classes responded to the test (i.e., the response rate was higher than 95%). Among their responses, about 8% of their responses were missing (students either left the answer blank or answered "I don't know"). Each of the items had more than 100 responses except for four elementary items that had slightly less than 100 responses per item.

Table 2. Information about the Schools that the Data Were Collected From

Schools	Great school overall rating	White student	Teacher	Number of tests collected
Elementary school 1	7	95%	Teacher 1	113
Elementary school 2	3	81%	Teacher 2	22
Elementary school 3	6	89%	Teacher 3&4	181
Middle school 1	6	89%	Teacher 5	253
Middle school 2	6	94%	Teacher 6	186
Middle school 3	NA	97%	Teacher 7	83
Middle school 4	7	93%	Teacher 8	205
High school 1	9	96%	Teacher 9	206
High school 2	10	98%	Teacher 10	115
High school 3	7	88%	Teacher 11&12	136

2.2. The Assessment

The assessment was administered to elementary, middle and high school students. The items administered at each grade level were selected to be appropriate for students at that grade level. At each grade level, a student typically took 10 to 12 items, including two or three two-tier items with the remaining being CR items. Approximately 25% of the items were anchor items that were used across all three-grade bands[1]. Though some anchor items are two-tier items, only the scores from the CR part were used in the linking process because of the small numbers of the OMC and MTF items in the assessment. The entire data set was calibrated through a concurrent calibration because there were common anchor items across grade levels.

The assessment has 42 items across three grade levels developed in 2009[2]. It includes 6 OMC + CR items, 10 MTF + CR items and 26 CR items[3]. Each of the 16 two-tier items (6

[1] Twenty percent of the total number of test items is recommended as the minimum number of anchor items in common-item linking (Angoff, 1971).

[2] A refined version of this assessment has been developed and has been used in a subsequent study (see Authors (2015) for details).

[3] A limitation of our two-tier items is the dependence between the first and second tier of the item which may confound interpretation of the selected explanation. However, in addition to CR questions in the two-tier items, we have 25 independent CR items that can be used to compare students' performance on different item formats without confounding factors.

OMC+CR and 10 MTF +CR items) was re-coded into two items, with each tier as a single item. This results in 42 items recoded into 58 items. It is worth noting that since we have small numbers of OMC and MTF items in this study, the generalizability of findings about these item formats need to be tested in future studies with more OMC and MTF items.

2.3. Scoring Process

The OMC responses were coded into learning progression levels according to the level of understanding that the choices represent. Students' responses to the MTF items, which included a string of T or F responses, were also coded into scores based on the number of correct choices made by the students. The responses to the CR items and the CR part of the two-tier items were coded using scoring rubrics aligned with the learning progression levels. Students' responses were coded by nine experienced raters with expertise in science education or educational measurement. Ten percent of the responses were double coded. The percentage of exact agreement between the first and second raters was 80% or higher for all items. Discrepancies in coding were discussed and final agreements were reached for each response.

2.4. Data Analysis

The data analysis process included two parts. First, we investigated how well the CR format classified students into learning progression levels using IRT based approaches. A partial credit model (PCM) was applied to model the scores of the CR responses. PCM is one of the most commonly used IRT models for polytomous items (Masters, 1982). In this study, the CR items and the CR part of the two-tier items had two or more score categories, and a higher score required accomplishing more of the desired task. Therefore, it was appropriate to use PCM to model the CR scores. For PCM, the probability of student *j* being graded into level *k* for item *i* is given by:

$$P(u_{ij} = k|\theta_j) = \frac{exp\sum_{q=0}^{k}(\theta_j - \delta_{qi})}{\sum_{k=0}^{m_i} exp\sum_{q=0}^{k}(\theta_j - \delta_{qi})}$$

where u_{ij} is the observed score of person j on item i,

m_i is the maximum score on item I,

θ_j is the proficiency of person j,

and δ_{qi} is the step parameter for the q^{th} score category for item i.

We first checked fit of PCM to the CR scores to see whether this model is appropriate for analyzing CR items. The results showed the model fit the CR scoring data well. When an item fits well with the model, the acceptable range of mean-squares (MNSQs) of the item is between 0.6 and 1.4 for polytomous items and the associated t statistics is within the range from -2 to 2 (Wu, Adams, Wilson, & Haldane, 2007; Wright, Linacre, Gustafsson, & Martin-Loff, 1994). The MNSQs and t-statistics for all items except one were within the acceptable ranges, which indicated our selection of model was appropriate. Table 3 presents the item fit statistics and difficulty parameter estimates of all the CR items.

Table 3. PCM Fit Results

ID	Item	Item Difficulty	ERROR	WEIGHTED FIT		
				MNSQ	CI	T
1	ACRON	0.75	0.05	1.08	(0.89, 1.11)	1.4
2	AIREV	-0.50	0.05	1.00	(0.81, 1.19)	0.1
3	AIRBO	1.01	0.05	0.93	(0.71, 1.29)	-0.5
4	ANIMW	-1.50	0.06	1.15	(0.74, 1.26)	1.1
5	APPLR	-0.34	0.05	0.91	(0.84, 1.16)	-1.1
6	BODYE	-0.01	0.05	1.07	(0.84, 1.16)	0.8
7	BODYH	-0.72	0.05	1.27	(0.78, 1.22)	2.2
8	BREAD	-0.21	0.04	1.18	(0.87, 1.13)	2.7
9	MATCHEL	-0.75	0.05	0.90	(0.81, 1.19)	-1.1
10	MATCHMA	0.09	0.05	1.05	(0.88, 1.12)	0.9
11	MATCHMB	0.12	0.05	0.97	(0.87, 1.13)	-0.5
12	CARBO	1.10	0.05	1.11	(0.75, 1.25)	0.8
13	CARPA	0.67	0.05	0.91	(0.76, 1.24)	-0.8
14	CARGA	-0.83	0.05	0.96	(0.85, 1.15)	-0.5
15	CONNL	-0.67	0.06	1.01	(0.80, 1.20)	0.1
16	CUTTR	-0.20	0.06	1.22	(0.65, 1.35)	1.2
17	DEERW	1.09	0.06	1.00	(0.74, 1.26)	0.0
18	DIFEV	0.02	0.05	1.05	(0.87, 1.13)	0.8
19	EATAP	-0.03	0.05	0.98	(0.88, 1.12)	-0.3
20	EATBR	0.43	0.05	1.09	(0.85, 1.15)	1.1
21	ECOSP	0.66	0.05	1.11	(0.81, 1.19)	1.1
22	ENERP	-0.09	0.04	1.04	(0.90, 1.10)	0.8
23	ENPLN	-0.23	0.04	1.04	(0.89, 1.11)	0.7
24	GIRLAB	-0.27	0.06	0.97	(0.71, 1.29)	-0.2
25	GIRLC	-0.74	0.06	0.80	(0.76, 1.24)	-1.7
26	GLOBM	-0.48	0.05	0.92	(0.80, 1.20)	-0.8
27	GLOBH	-0.88	0.05	0.97	(0.77, 1.23)	-0.2
28	GLUEG	-1.07	0.05	0.92	(0.83, 1.17)	-0.9
29	GRAND	0.85	0.06	1.03	(0.70, 1.30)	0.3
30	GRANP	0.06	0.06	0.93	(0.72, 1.28)	-0.5
31	GROWT	0.92	0.05	0.92	(0.85, 1.15)	-1.0
32	INFAN	0.74	0.05	0.95	(0.89, 1.11)	-0.8
33	KLGSE	0.60	0.05	1.05	(0.77, 1.23)	0.5
34	LAMPE	-0.41	0.05	1.17	(0.85, 1.15)	2.1
35	OCTAM	0.48	0.05	0.90	(0.79, 1.21)	-0.9
36	PLANG	0.82	0.05	0.92	(0.87, 1.13)	-1.2
37	THINT	-0.31	0.04	0.82	(0.89, 1.11)	-3.6
38	TREDEAB	0.17	0.04	0.93	(0.89, 1.11)	-1.3
39	TREDEC	-0.06	0.04	0.96	(0.88, 1.12)	-0.6
40	TROPRA	-1.08	0.05	1.01	(0.89, 1.11)	0.2
41	WAXBUR	1.49	0.06	0.98	(0.71, 1.29)	-0.1
42	WTLOSS	-0.682	0.32	1.15	(0.91, 1.09)	3.0

Note: ERROR = the error of the item difficulty estimates; MNSQ = the mean residual square between what is observed and what is expected; CI = the confidence interval of the MNSQ; T = the t-statistics that used to indicate the fitness of the item to the model. T value with * indicates that the item does not fit well with PCM.

Results from the IRT analysis such as the item fit indices, item difficulty and step parameter were used to evaluate the effectiveness of the CR items in distinguishing students among levels and items that did not converge or show poor IRT fit were reviewed. ConQuest (Wu, Adams, & Wilson, 1998) was used to estimate the item and person ability parameters. This software was designed to analyze data based on the multidimensional random coefficients multinomial logit model (MRCMLM; Adams, Wilson, & Wang, 1997) and PCM is a special case of MRCMLM.

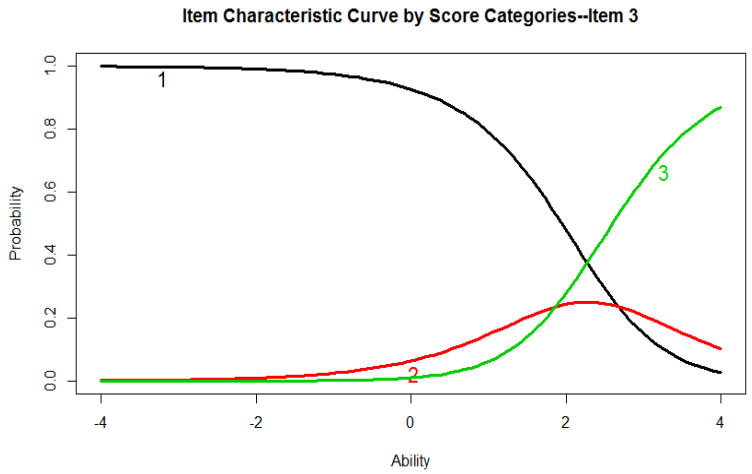
Second, we investigated how well the OMC and MTF formats classified students' responses into learning progression levels by comparing the level of a student's OMC or MTF responses to the level of his/her CR responses. Here, we used the student's CR responses to gauge the effectiveness of OMC and MTF formats because if the OMC and MTF items can accurately predict the level of students' CR responses, they can be used as an alternative to the CR items to reduce administration time and scoring cost. We did not conduct IRT analysis for the OMC and MTF item formats because there were relatively fewer OMC and MTF questions in the item pool. One student typically only answered two or three OMC or MTF questions, which causes the item and person ability estimates to be unstable when conducting IRT analysis.

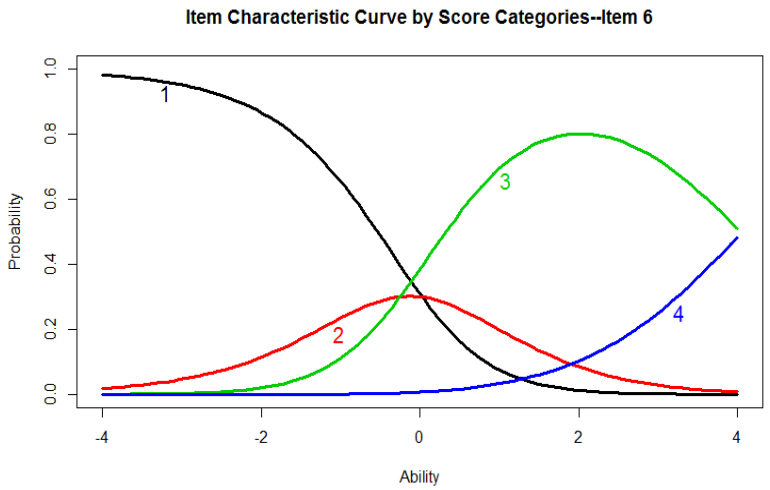
3. RESULTS

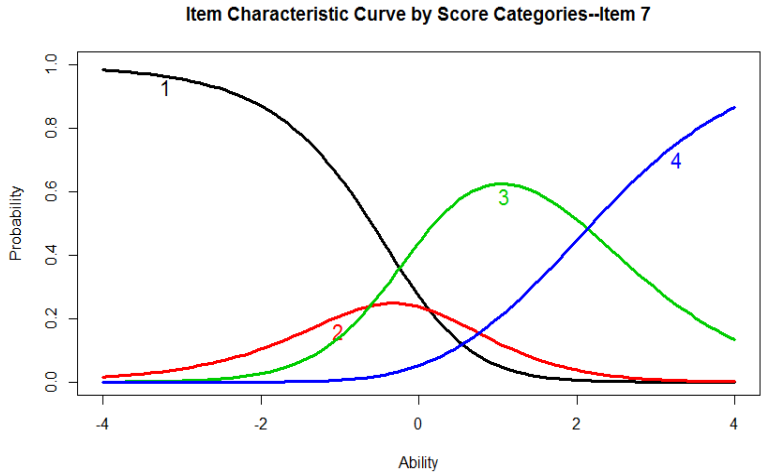
3.1. The Effectiveness of the CR, OMC, and MTF Item Formats in Classifying Students among Levels

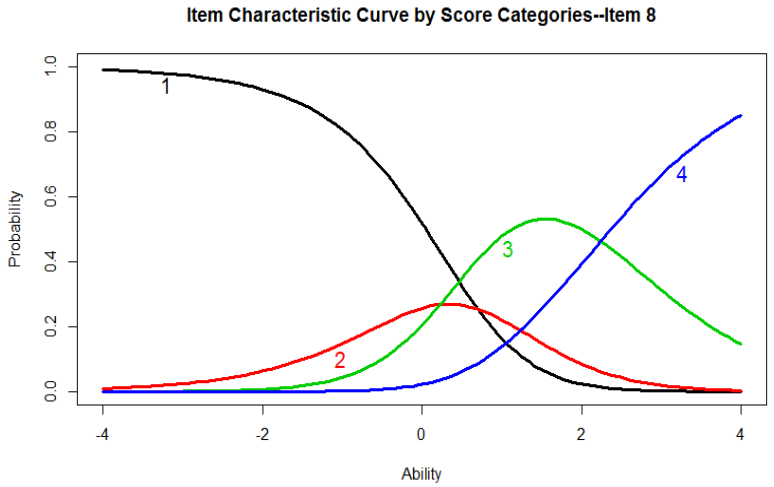
CR format. We analyzed the step parameters (i.e. Thurstonian thresholds) of score categories estimated from the PCM model. The step parameter d_1 is the ability level at which the student has the same probability of providing a level 1 and a level 2 answer. When the student's ability level increases, he/she has a larger probability of providing a level 2 than a level 1 answer. The step parameter d_2 is the ability level at which the student has the same probability of providing a level 2 and a level 3 answer, and so on. Theoretically, the step parameters should be in the correct order ($d_1<d_2<d_3$) because with the increase of ability, students are more likely to be at higher levels. If the step parameters are not in the correct order, it might indicate that the item was not discriminative at certain learning progression levels or the item classified students inaccurately.

The results indicates that most of the CR items are effective for differentiating students among learning progression levels. Among all 42 items (26 CR items and 16 CR questions from the two-tier items), the step parameters of 35 items were in the correct order. The step parameters of 7 items were not in the correct order. This may suggest that there were too many or too few responses at a particular level compared to the proportion of responses at that level from other items. So these items did not accurately classify students at particular levels. Figure 1 presents the item characteristic curves by categories for these eight items with step parameters in the wrong order.


Item Characteristic Curve by Score Categories--Item 3
Probability
Ability
1
2
3


Item Characteristic Curve by Score Categories--Item 6
Probability
Ability
1
2
3
4


Item Characteristic Curve by Score Categories--Item 7
Probability
Ability
1
2
3
4


Item Characteristic Curve by Score Categories--Item 8
Probability
Ability
1
2
3
4

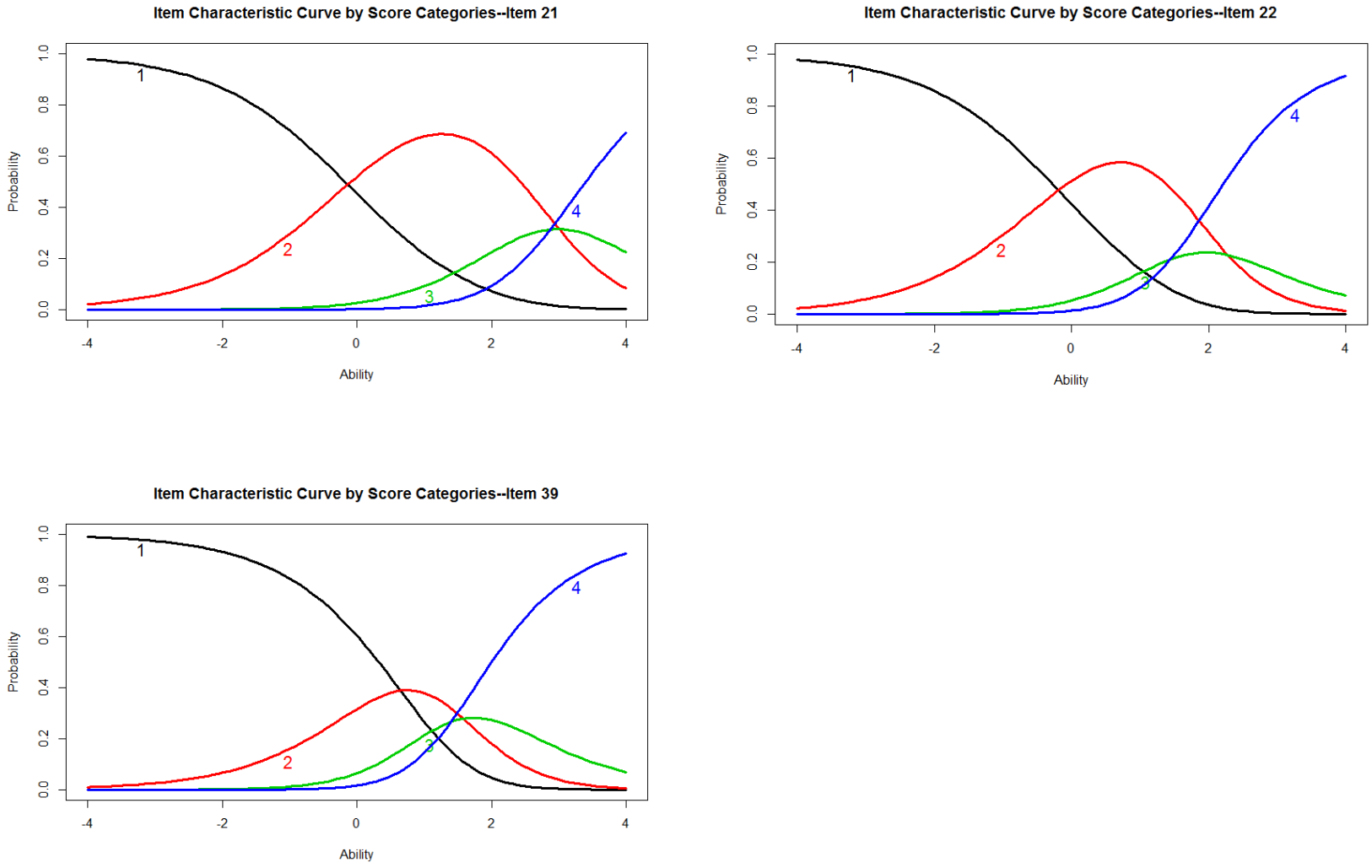


Figure 1. The Characteristic Curves by Category of Seven CR Items

We found that CR items that asked for macroscopic explanations of phenomena were not able to elicit level 4 responses and items that required microscopic explanations of phenomena were not able to elicit level 1 responses. For example, the OCTAMOLE item and the CARGAS item both assessed the concept of the combustion of gasoline. Both items ask students what happens to gasoline when the gasoline tank of a car gets empty after running. The OCTAMOLE item tells students that gasoline is mostly a mixture of hydrocarbons such as octane (C_8H_{18}) and asks them what happens to gasoline. In contrast, the CARGAS item does not provide the chemical identify of gasoline. The difficulty of the OCTMOLE item was 0.5 (i.e. above average) and the difficulty of the CARGAS item was -0.8 (i.e., below average). The item information curves are presented in Figure 2. The information for OCTAMOLE peaked in the high ability range and the CARGAS had a relatively flat information curve over a wider range of abilities. Therefore, depending on the students who take the item, the same question can be asked in different scales (e.g. microscopic scale, macroscopic scale) to measure students more precisely.

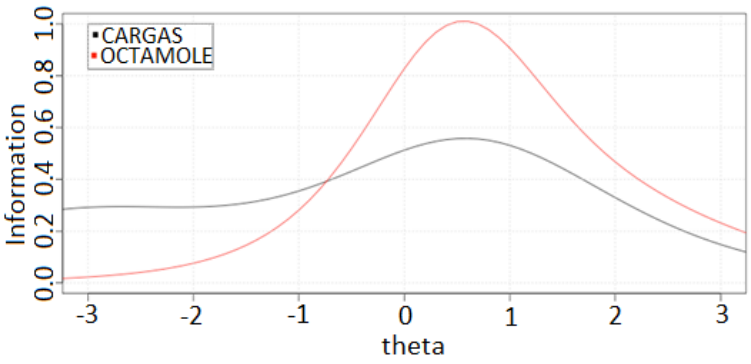


Figure 2. The Item Information Curves of the CARGAS and the OCTAMOLE Items

OMC options. The level of a student based on his/her OMC and MTF responses was compared to his/her level based on CR responses. The CR format is used to gauge the effectiveness of OMC and MTF formats because CR responses usually have richer information that provide a better measure of students' 'true' learning progression levels. The levels of the OMC options are determined based on the same rubric as the CR responses and so are comparable to the levels of CR responses. Three columns in the middle of Table 4 show the percentage of responses for which the OMC level correctly predicts, over predicts, or under predicts the CR level. This result shows that in some cases OMC level can predict the CR level, but there are cases in which the level of a student's OMC response is higher or lower than the level of his/her CR response.

A student's OMC level may over or under-predict his/her CR level because the OMC choices are associated with a restricted range of learning progression levels. Table 4 lists the range of achievement levels of each OMC item. Some of the items do not have a level 4 option or a level 1 option, which make them unable to measure students precisely at the two ends of the achievement level scale. For instance, the MATCH item has four OMC options at level 1 to 3. 34.6% of students received higher levels on the paired CR item than their levels on the OMC item.

Table 4. The Proportions of Accurate, Over and Under-Prediction of OMC Questions

Item number	Item name	Level range of OMC choices	Paired CR			Average of all other CR items		
			Correct prediction (%)	Over-prediction (%)	Under-prediction (%)	Correct prediction (%)	Over-prediction (%)	Under-prediction (%)
1	DEER	1~3	63.7	10.3	26.0	60.9	13.5	25.6
2	ACORN	1~3	56.7	32.7	10.6	43.5	28.4	28.1
3	WTLOSS	1~3	51.7	17.7	30.6	53.3	17.6	29.1
4	TROPRAIN	2~4	51.3	39.8	8.9	51.2	26.1	22.8
5	MATCH	1~3	44.4	21.0	34.6	40.7	46.3	13.0
6	BREAD	2~4	31.3	56.4	12.3	48.3	37.8	13.9

Therefore, in this case, the OMC option did not predict the level of the student's CR response very well.

Item Match: When a match burns, the released energy

A) comes mainly from the match. (Level 3)
B) comes mainly from the air. (Level 2)
C) is created by the fire. (Level 1)
D) comes from the energy that you used to strike the match. (Level 2)

Please explain your answer.

Another cause of the mismatch of the same students' OMC and CR levels is that the OMC and the CR questions do not assess the same ability though they target the same concept. The OMC item asks students to identify the best answer but the CR item asks students to explain why they think the choice they made is the best answer. Students' OMC levels sometimes over-predict their CR levels because they do a better job identifying the best choice than explaining their choice. One example is the ACORN item described previously that asks students to identify source of the weight of the oak tree. Students were able to choose the correct choice that the weight of the oak tree came from CO_2 and water but could not explain how CO_2 and water contributed to weight gain. This pattern was common for other OMC + CR items. In particular, far more students were able to identify Level 4 responses in OMC items than were able to produce Level 4 responses to CR items.

The dependence between the OMC and the CR questions in the two tier items may inflate the prediction of OMC format. To avoid this, we analyzed how well OMC questions predict the learning progression levels measured by all the other CR items. The last three columns in Table 4 shows the percentage of responses that the level of a student's OMC response correctly predicts, over, or under-predicts the average of the levels of his/her responses to all the other CR items. This result shows the OMC level can also predict the level of students' responses to all the CR items to some extent. All the findings about OMC items needs to be verified in future studies using more OMC items since we only have a small number of OMC items (i.e. six items) in this study.

MTF format. Similar to OMC format, we evaluated how well students' *T* or *F* choices predicted the level of their CR responses. The empirical data suggested that the response string to the set of True or False questions did not clearly associate with the learning progression levels. For example, a MTF question with four T or F questions can have 16 different combinations of True or False choices. There are no clear ways to associate these 16 different response strings with the four achievement levels. Thus, we analyzed the number of

International Journal of Assessment Tools in Education: Vol. 3, Issue 2 (2016) pp. 101-122

correct True or False choices to simplify our analysis of the MTF format. For instance, if a student made one correct choice, his/her score of that MTF question is 1; but if he/she made four correct choices, his/her score of that question is 4. Two patterns of the MTF format were found.

First, the number of correct choices of most MTF questions had only a weak to moderate correlation with students' CR level. Table 5 presents the Kendall's tau correlation between students' MTF scores (number of correct choices), their levels on the paired CR questions, and the average of their levels on all the other CR items. This finding suggests the number of correct choices of the MTF items can provide some information about a student's level of his/her CR responses.

Table 5. Kendall's Tau Correlation between Students' Number of Correct Choices and Their Levels on the CR Items

	MTF items	Score Scale	N	Tau between No. of correct choices and the level of paired CR question	Tau between No. of correct choices and the average level of all other CR items
1	GLOBWARM_H	0~5	132	0.48	0.37
2	ENERPEOP	0~7	469	0.42	0.35
3	OCTAMOLE	0~6	147	0.39	0.36
4	GLOBWARM_M	0~4	164	0.26	0.32
5	BODYTEMP	0~4	189	0.18	0.26
6	ENERPLNT	0~5	508	0.16	0.23
7	AIREVENT	0~4	189	0.16	0.18
8	THINGTREE	0~4	585	0.12	0.02
9	INFANT	0~4	450	0.02	0.12
10	ANIMWINTER	0~5	78	0.02	0.24

Second, some true/false questions work better to predict the learning progression level of a student's CR responses than others. We conducted independent sample *t*-test to test the mean difference of students' average CR levels between two groups of students: the group of students who selected "True" and the group of students who selected "False" for each True or False question. We found that for some True or False questions, students who selected "True" were at significantly different levels in terms of their CR responses from those who selected "False," which suggests the questions were effective in differentiating students. Table 6 list the numbers of the students who made True choice and that of the students who made False choice, the mean and standard deviation of the average CR levels of each group, and the *t*-test results. For example, in the ENERPLNT item presented previously, the "water", "air", and "nutrients in soil" options are most effective to detect students' differences. For three options, water, air and nutrient, students who circled "No" gave significantly higher-level CR responses than those who selected "Yes" (water: p-value <0.001, air: p-value<0.001, nutrient: p-value<0.001). For the other two options, "light" and "plants make their own energy," the differences between students who selected "Yes" and students who selected "No" were not significant (light: p-value=0.276, own energy: p-value=0.606). This means that these two options were less effective in differentiating students. It's worth noting that the findings about MTF items need to be verified in future studies using more MTF items since we only have a small number of MTF items in this study.

Table 6. Independent T-test Results About the Mean Different Between Students Who Selected True and Students Who Selected False for Each T or F Question.

	T or F questions	No. of True	No. of False	Mean of True	Mean of False	SE of True	SE of False	Sig.
1. GLOBWARM_H	Q1	126	9	2.40	1.89	0.55	0.33	.001
	Q2	110	24	2.43	2.08	0.57	0.41	.001
	Q3	74	62	2.47	2.23	0.58	0.49	.008
	Q4	104	30	2.36	2.37	0.57	0.49	.925
	Q5	37	98	2.65	2.24	0.54	0.52	.000
2. ENERPEOP	Q1	723	174	2.11	2.45	0.42	0.51	.000
	Q2	877	21	2.18	2.33	0.46	0.48	.124
	Q3	742	152	2.18	2.22	0.43	0.57	.324
	Q4	466	416	2.05	2.33	0.39	0.50	.000
	Q5	455	427	2.14	2.23	0.43	0.49	.003
	Q6	114	367	2.11	2.34	0.31	0.49	.000
	Q7	764	127	2.14	2.45	0.44	0.50	.000
3. OCTAMOLE	Q1	133	17	2.22	2.06	0.53	0.24	.039
	Q2	118	31	2.19	2.26	0.47	0.63	.558
	Q3	127	22	2.16	2.45	0.43	0.80	.103
	Q4	48	100	2.02	2.28	0.33	0.55	.000
	Q5	128	21	2.13	2.62	0.42	0.74	.008
	Q6	81	67	2.35	2.01	0.57	0.33	.000
4. GLOBWARM_M	Q1	161	24	2.04	1.79	0.46	0.59	.060
	Q2	113	67	2.14	1.78	0.48	0.42	.000
	Q3	67	112	2.10	1.94	0.53	0.45	.026
	Q4	137	43	2.02	1.98	0.45	0.60	.649
5. BODYTEMP	Q1	72	122	1.81	2.02	0.43	0.37	.000
	Q2	157	40	1.96	1.88	0.39	0.46	.281
	Q3	87	105	1.83	2.04	0.41	0.39	.000
	Q4	172	26	1.96	1.88	0.41	0.33	.376
6. ENERPLNT	Q1	454	65	2.11	2.58	0.43	0.58	.000
	Q2	498	22	2.16	2.32	0.47	0.65	.276
	Q3	319	197	2.10	2.27	0.44	0.51	.000
	Q4	431	87	2.10	2.52	0.43	0.55	.000
	Q5	233	278	2.15	2.18	0.48	0.47	.606
7. AIREVENT	Q1	160	38	2.01	1.87	0.44	0.34	.074
	Q2	177	20	1.99	1.85	0.41	0.59	.296
	Q3	156	41	2.03	1.78	0.42	0.42	.001
	Q4	59	131	1.97	2.00	0.45	0.41	.611
8. THINHTREE	Q1	582	16	2.05	1.69	0.52	0.70	.055
	Q2	512	83	2.00	2.29	0.50	0.57	.000
	Q3	576	17	2.04	2.18	0.53	0.53	.298
	Q4	407	186	2.10	1.92	0.56	0.44	.000
9. INFANT	Q1	165	294	1.92	2.02	0.52	0.54	.000
	Q2	431	33	1.97	2.15	0.54	0.51	.044
	Q3	385	78	1.96	2.08	0.55	0.45	.055
	Q4	461	2	1.98	2.00	0.54	0.00	.076
10. ANIMWINT	Q1	49	31	1.84	1.87	0.43	0.43	.727
	Q2	24	56	1.79	1.88	0.42	0.43	.424
	Q3	34	46	1.68	1.98	0.48	0.33	.001
	Q4	20	60	1.70	1.90	0.47	0.40	.067
	Q5	37	42	1.81	1.90	0.52	0.30	.336

International Journal of Assessment Tools in Education: Vol. 3, Issue 2 (2016) pp. 101-122

3.2. How to Design a Test with a Mixture of CR, OMC and MTF formats to Measure Students' Learning Progression Levels?

These findings provide some suggestions for designing a test with a mixture of item formats to measure students' learning progression levels. A test composed of items in different formats can utilize the advantages of each format to measure a learning progression more precisely. In this section, we first discuss the design of CR items in a learning progression based test and then discuss the design of OMC and MTF items.

For learning progression based assessment, ideally, a student will get the same learning progression level across all the items. This means that the item step parameters of the CR items for the same score category need to be similar across items. Figure 3 presents the distribution of item step parameters of our CR items. It can be seen that the parameters of the same step across items vary a bit. If the items and rubrics are well designed and the scorings are reliable, then the variance of the step parameters, d_1, d_2 and d_3 should be small. The items that had d_1, d_2 and d_3 deviating far from the mean values suggest that either the item or the scoring is not appropriate. For example, on the "histogram of d_1 graph", the d_1 of some items are much smaller than the others. These items will not do a good job discriminating at the lowest level-level 1.

Including OMC and MTF formats in a test can establish the reliability of the test because a larger number of items can be administrated. Our prior analysis suggested that some OMC questions did not predict the level of students' CR responses very well and some True or False questions did not differentiate students effectively among levels. To use OMC and MTF items in a test, we need to design and use these items carefully. OMC and MTF questions are often only associated with a restricted range of learning progression levels; meaning that we can use them to distinguish students among some levels better than others. An OMC or MTF item may only provide good information about the distinction between two levels (e.g. between "level 4" and "below level 4"), but not about the distinction among levels 1, 2, and 3. In this case, the OMC or MTF item can be treated as a dichotomous item that the responses are recoded into either "1," the best choice, or "0," all the other choices. However, in this case, these OMC items do not work better than traditional multiple choice items that only provide distinction between correct or incorrect answers.

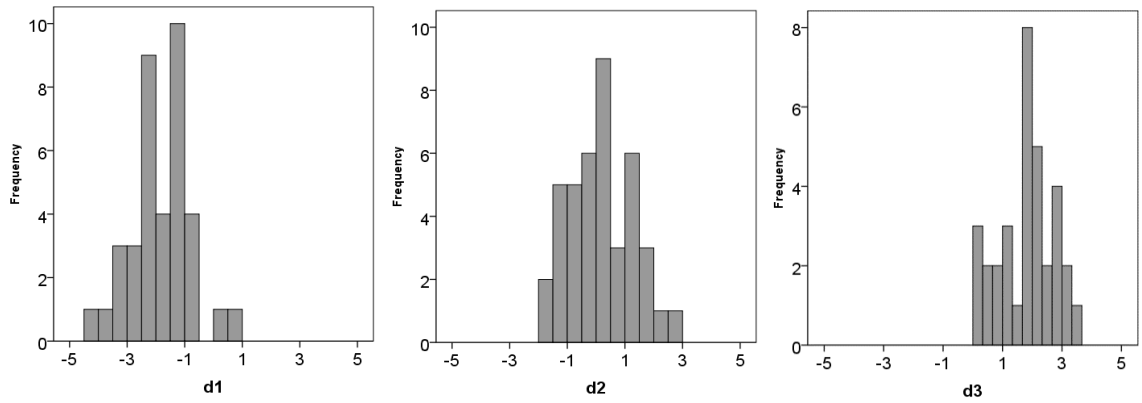


Figure 3. Distribution of Step Threshold Parameters

Note: d1, d2 and d3 are cutting point on the ability scale between score 1 and 2, 2 and 3, and 3 and 4. For example, if a student's ability is d_1, then he/she will have 50% of the chance to get either a Level 1 or a Level 2 score. If his/her ability increases, the probability of getting a level 2 will be higher than the probability of getting a level 1.

For example, the OMC item (TROPRAIN) presented below only distinguish students who made the correct answer "b", and those who made the incorrect answers "a", "c" and "d". For the students who made the incorrect choices, it's difficult to determine their learning progression levels based on their choices. So this item can be treated as a dichotomous item that only valid to differentiate students between two levels.

TROPRAIN

A tropical rainforest is an example of an ecosystem. Which of the following statements about matter and energy in a tropical rainforest is the most accurate? Please choose ONE answer that you think is best.

a. Energy is recycled, but matter is not recycled. (Below level 4)
b. Matter is recycled, but energy is not recycled. (Level 4)
c. Both matter and energy are recycled. (Below level 4)
d. Both matter and energy are not recycled. (Below level 4)

4. IMPLICATIONS

4.1. Broader Implications in Developing Learning Progression-Based Assessments

This study provided an example of a test with a mixture of different item formats that measured learning progression levels. There are three advantages to using different item formats in a test. First, a combination of different formats may measure all aspects of given construct. Consistent with previous studies (Yao & Boughton, 2009; Lee, Liu, & Linn, 2011), results from this study suggest that items in different formats might assess slightly different aspects of the learning progression. For example, even when assessing the same phenomenon, OMC items focus on students' ability to identify the best choice but CR items mainly assess students' ability to organize and synthesize their knowledge to solve problems. Thus, a combination of different formats may provide more comprehensive information about students' ability.

Second, both the CR items and the selected choice items (e.g., OMC, MTF) may not measure students' learning progression levels precisely at all levels. For example, MTF items may only work well to differentiate students at level 4 who often made all correct choices but not effective in differentiating students at the other levels. Therefore, a combination of different types of items that are effective at different learning progression levels can reduce measurement error at all the levels.

Finally, a combination of different item formats may facilitate the use of computerized adaptive testing (CAT) to measure students' learning progression levels precisely over a wide range of ability levels. After getting a stable estimate about students' achievement levels based on selected-choice items, selected CR items that target at corresponding levels can be administrated to collect more detailed information and make fine-grained distinctions. Measure students' learning progression levels using CAT can provide precise scores for most test-takers and save test administration time and scoring effort.

4.2. Implications for Developing Items in OMC, MTF and CR Formats

This study provides some suggestions for best using each of the formats. First, one challenge with developing learning progression-based CR items is that it is difficult to write items that provide opportunities for students to respond at multiple levels of a learning progression (Anderson, Alonzo, Smith, & Wilson, 2007). The results of this study also suggest some CR items can only elicit responses at particular levels rather than all levels. Items proposed at microscopic scale are generally only discriminative for level 2 and above.

International Journal of Assessment Tools in Education: Vol. 3, Issue 2 (2016) pp. 101-122

For instance, an item asks examinees "what happens to the atoms in amylose molecules as the potato decays" can only elicit responses that are level 2 and above. These items need to be used appropriately so they are discriminative for the examinees who take the items (i.e., macroscopic items for lower level students and microscopic items for higher level students).

Second, the choices of OMC items are associated with a restricted range of learning progression levels, which may make students' OMC levels over or under-predict their CR levels. Items that do not have a level 4 option or a level 1 option cannot measure students precisely at the two ends of the achievement level scale, thus would not be appropriate for less sophisticated or very sophisticated students. This is similar to the findings from previous studies that OMC items provide less precise measures for high and low ends of the ability distribution than CR items (Ercikan et al., 1998; Wilson & Wang, 1995). Including OMC options at all (or most) learning progression levels might be ideal; however, it is difficult to write OMC options at higher achievement levels without using "science-y" terminologies that indicate the highest-level response (Alonzo & Gotwals, 2012) and it is difficult to design options to measure students in the low ability range precisely since answering OMC item involves guessing and examinees in the low ability range tend to guess most (Taleto-Miller, Han, & Guo, 2011). In order to reduce measurement errors due to the low discrimination of the OMC format at the high end or low end, we may choose to only use OMC items to measure students at the middle learning progression levels where the OMC options are more discriminative.

Third, one finding about the MTF items is that some true or false options are more effective than others in differentiating students. In our case, choices of water, air, and nutrients *NOT* being energy sources for plants were effective options to distinguish between high level and lower level students. However, just choosing sunlight as an energy source or choosing that plants make their own energy did *NOT* discriminate between higher level and lower level students. This may indicate that being able to accurately falsify certain combinations of alternative ideas is more telling of deep understandings than others. However, in order to design discriminative MTF options, more research is needed to find the most effective options and most efficient combinations of options to differentiate students. Results from this study also suggest the number of correct choices provides some information about students' average CR level. There are weak to moderate correlations between students' number of correct choices and the average of their CR levels. However, to make MTF items more predictive of students' CR levels, the MTF options need to be more discriminative as discussed previously.

5. LIMITATIONS AND FUTURE WORK

Some problems limit the validity or generalizability of the findings from this study and suggest directions for future work. First, this study used data from 6 OMC and 10 MTF items. So the findings about the OMC and MTF formats were based on data from relatively small numbers of items. These findings need to be verified in the future with data from more OMC and MTF items. We did not conduct IRT analysis for the OMC and MTF items because the data are sparse which will lower the precision of item parameter estimates. When complete data is available in future studies, it's worth applying IRT models to fit these items to examine the quality and the characteristics of these items (e.g. how discriminative is an OMC item, and at which level is the item most discriminative).

Consistent with previous studies (Lee, Liu, & Linn, 2011;Yao & Boughton, 2009), we noticed that the selected choice items (i.e. OMC; MTF) and the CR question do not assess

exactly the same ability though they target the same construct. Thus, though our results suggest students' responses to the OMC and MTF items can predict their CR responses to some extent, these item formats cannot be used interchangeably without careful consideration of the construct being measured by different item formats.

6. REFERENCES

Adams, R.J., Wilson, M., & Wang, W-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1-23.

Alonzo, A.C., & Gotwals, A. G. (2012). *Learning progressions in science.* Rotterdam, The Netherlands: Sense Publishers.

Alonzo, A.C., & Steedle, J. T. (2008). *Developing and assessing a force and motion learning progression.* Published online in Wiley InterScience.

Anderson, C.W., Alonzo, A. C., Smith, C., & Wilson, M. (2007, August). *NAEP pilot learning progression framework.* Report to the National Assessment Governing Board.

Angoff, W.H. (1971). *Scales, norms, and equivalent scores.* In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Berlak, H. (1992). The need for a new science of assessment. In H. Berlak et al. (Eds.) *Toward a new science of educational testing and assessment* (pp. 1-22). Albany: State University of New York Press.

Briggs, D.C. & Alonzo, A.C. (2012). *The psychometric modelling of ordered multiple-choice item responses for diagnostic assessment with a learning progression.* In A.C. Alonzo & A.W. Gotwals (eds). *Learning progressions in science: Current challenges and future directions.* Rotterdam, The Netherlands: Sense Publishers.

Briggs, D.C., Alonzo, A.C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*, 33 – 63.

Catley, K., Lehrer R., & Reiser, B. (2004). *Tracing a prospective learning progression for developing understanding of evolution,* Paper Commissioned by the National Academies Committee on Test Design for K–12 Science Achievement, Washington, DC: National Academy of Science, 67.

Chen, J. & Anderson, C.W. (2015). Comparing American and Chinese K-12 students' learning progression on carbon cycling in socio-ecological systems. *Science Education International.* 27(4), 439-462.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009, May). *Learning progressions in science: An evidence based approach to reform* (CPRE Research Report #RR-63). Philadelphia, PA: Consortium for Policy Research in Education.

Doherty, J., Draney, K., Shin, H., Kim, J., & Anderson, C. W. (2015). Validation of a learning progression-based monitoring assessment. Manuscript submitted for publication.

Downing, S. M., & Yudkowsky, R. (2009). *Assessment in Health Professions Education*. New York, NY Routledge.

Dunham, M. L. (2007) An investigation of the multiple true-false item for nursing licensure and potential sources of construct-irrelevant difficulty. http://proquest.umi.com/pqdlink?did=1232396481&Fmt=7&clientId=79356&RQT=309&VName=PQD

Embretson, S. E. (1996). Item response theory models and inferential bias in multiple group comparisons. *Applied Psychological Measurement, 20,* 201-212.

International Journal of Assessment Tools in Education: Vol. 3, Issue 2 (2016) pp. 101-122

Ercikan, K., Schwarz, R.D., Julian, M. W., Burket, G.R., Weber, M.M. & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement, 35*, 137-154.

Flowers, K., Bolton, C., Brindle, N. (2008). Chance guessing in a forced-choice recognition task and the detection of malingering. *In: Neuropsychology, 22* (2), 273-277

Frisbie, D. A. (1992). The multiple true-false item format: A status review. *Educational Measurement: Issues and Practice,*11(4), 21–26.

Jin, H., & Anderson, C. W. (2012). A learning progression for energy in socio-ecological systems. *Journal of Research in Science Teaching, 49*(9), 1149–1180.

Lee, H.S., Liu, O.L. & Linn, M.C. (2011). Validating Measurement of Knowledge Integration Science Using Multiple-Choice and Explanation Items. *Applied Measurement in Education, 24*(2), 115-136.

Liu, O.L., Lee, H-S., Hofstedder, C. & Linn, M.C. (2008). Assessing knowledge integration in science: Construct, measures and evidence. *Educational Assessment. 13*, 33-55.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-173.

Martinez, M. (1999). Cognition and the question of test item format. *Educational Psychologists, 34*, 207- 218.

Merritt, J. D., Krajcik, J. & Shwartz, Y. (2008). *Development of a learning progression for the particle model of matter.* In Proceedings of the 8th International Conference for the Learning Sciences (Vol. 2, pp. 75-81). Utrecht, The Netherlands: International Society of the Learning Sciences.

Mohan, L., Chen, J., & Anderson, C.W. (2009). Developing a multi-year learning progression for carbon cycling in socio-ecological systems. *Journal of Research in Science Teaching*. 46 (6), 675-698.

National Research Council. (2006). *Systems for state science assessment.* Washington, DC: The National Academies Press.

National Research Council. (2007). *Taking science to school*. Washington, DC: The National Academies Press.

National Research Council. (2014). *Developing Assessments for the Next Generation Science Standards. Committee on Developing Assessments of Science Proficiency in K-12. Board on Testing and Assessment and Board on Science Education,* J.W. Pellegrino, M.R. Wilson, J.A. Koenig, and A.S. Beatty, Editors. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Plummer, J.D. & Maynard, L. (2014). Building a learning progression for celestial motion: An exploration of students' reasoning about the seasons. *Journal of Research in Science Teaching, 51*(7), 902-929.

Rivet, A. & Kastens, K. (2012). Developing a construct-based assessment to examine students' analogical reasoning around physical models in earth science. *Journal of Research in Science Teaching, 49*(6), 713-743.

Salinas, I. (2009, June). *Learning progressions in science education: Two approaches for development.* Paper presented at the Learning Progressions in Science (LeaPS) Conference, Iowa City, IA. *Available from* http://www.education.uiowa.edu/projects/leaps/proceedings/

Schuwirth, L. & van der Vleuten, C. (2004). Different Written Assessment Methods: What can be said about their Strengths and Weaknesses? *Medical Education 38*,9: 974-979.

Songer, N.B. & Gotwals, A.W. (2012). Guiding explanation construction by children at the entry points of learning progressions. *Journal for Research in Science Teaching, 49*, 141-165.

Songer, N.B., Kelcey, B. & Gotwals, A.W. (2009). How and when does complex reasoning occur? Empirically driven development of a learning progression focused on complex reasoning in biodiversity. *Journal of Research in Science Teaching*. *46*(6): 610-631.

Smith, C.L., Wiser, M., Anderson, C.W., & Krajcik, J. (2006). Implications on research on children's learning for standards and assessment: A proposed learning for matter and the atomic molecular theory. *Measurement: Interdisciplinary Research & Perspective, 4*(1), 1-98.

Steedle, J.T. & Shavelson, R.J. (2009). Supporting valid interpretations of learning progression level diagnoses. *Journal of Research in Science Teaching, 46*(6), 699-715.

Talento-Miller, E., Han, K. & Guo, F. (2011). *Guess Again: The Effect of Correct Guesses on Scores in an Operational CAT Program.* (Graduate Management Admission Council research report. No. RR-11-04). http://www.gmac.com/~/media/Files/gmac/Research/research-report-series/guessagaintheeffectofcorrect.pdf

Thissen, D., & Steinberg, L. (1997). *A response model for multiple-choice items.* In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 52-65). New York: Springer-Verlag.

van der Linen, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6*(2), 103-118.

Wilson, M., & Wang, W.C. (1995). Complex composites: Issues that arise in combining different modes of assessment. *Applied Psychological Measurement, 19*(1), 51-71.

Wright, B.D., Linacre, J.M., Gustafsson, J.E. & Martin-Loff, P. (1994). Reasonable mean-square fit values. *Rasch Meas Trans 1994; 8*: 370.

Wu, M.L., Adams, R.J. & Wilson, M.R. (1998). *ACER Conquest: Generalised item response modelling software.* Melbourne: ACER Press.

Wu, M.L., Adams, R.J., Wilson, M. R. & Haldane, S. A. (2007). *ACER ConQuest Version 2.0: generalised item response modeling software.* Camberwell, Australia: Australia Council for Educational Research.

Yao, L., & Boughton, K.A. (2009). Multidimensional linking for tests with mixed item types. *Journal of Educational Measurement, 46* (2), 177–197.