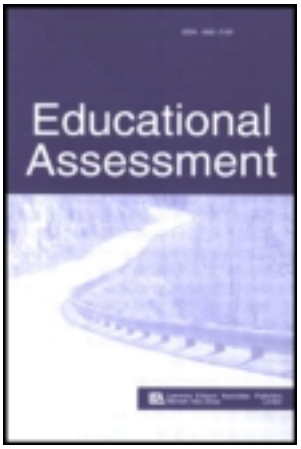

Educational Assessment
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/heda20

An Investigation of Explanation Multiple-Choice Items in Science Assessment

Ou Lydia Liu [a], Hee-Sun Lee [b] & Marcia C. Linn [b]
[a] Educational Testing Service
[b] University of California, Berkeley

Available online: 08 Sep 2011

To cite this article: Ou Lydia Liu, Hee-Sun Lee & Marcia C. Linn (2011): An Investigation of Explanation Multiple-Choice Items in Science Assessment, Educational Assessment, 16:3, 164-184

To link to this article: http://dx.doi.org/10.1080/10627197.2011.611702

Educational Assessment, 16:164–184, 2011

ISSN: 1062-7197 print/1532-6977 online
DOI: 10.1080/10627197.2011.611702

Routledge
Taylor & Francis Group

An Investigation of Explanation Multiple-Choice Items in Science Assessment

Ou Lydia Liu
Educational Testing Service

Hee-Sun Lee and Marcia C. Linn
University of California, Berkeley

Both multiple-choice and constructed-response items have known advantages and disadvantages in measuring scientific inquiry. In this article we explore the function of explanation multiple-choice (EMC) items and examine how EMC items differ from traditional multiple-choice and constructed-response items in measuring scientific reasoning. A group of 794 middle school students was randomly assigned to answer either constructed-response or EMC items following regular multiple-choice items. By applying a Rasch partial-credit analysis, we found that there is a consistent alignment between the EMC and multiple-choice items. Also, the EMC items are easier than the constructed-response items but are harder than most of the multiple-choice items. We discuss the potential value of the EMC items as a learning and diagnostic tool.

The science education standards call for students to develop coherent understanding of complex science topics (American Association for the Advancement of Science, 1993; National Research Council, 1996). Students with complex understanding should be able to know the principles that underlie science phenomena and be able to provide explanations for science phenomena using coherent evidence (Linn & Hsi, 2000; Linn, Lee, Tinker, Husic, & Chiu, 2006). However, the measurement of complex understanding is challenged by the lack of an ideal assessment format: Both multiple-choice (MC) and constructed-response (CR) items are widely used in measuring complex science understanding, yet they each have their known advantages and disadvantages. MC items are often criticized for focusing on recollection of scientific facts or straightforward applications of process skills rather than promoting standards-based coherent ideas (Clark & Linn, 2003; Heubert & Hauser, 1999; Shepard, 2000). CR items have the power to capture complex reasoning and student justification but are often challenged by high costs of administration and scoring and low reliabilities. To take advantage of both assessment formats,

Correspondence should be sent to Ou Lydia Liu, Foundational & Validity Center, Educational Testing Service, 660 Rosesale Road, Mailstop 16-R, Princeton, NJ 08534. E-mail: lliu@ets.org

researchers seek powerful alternatives to enhance the reasoning function of MC items and reduce the cost of CR items (Briggs, Alonzo, Schwab, & Wilson, 2006; Sadler, 1998). In the following section, we (a) review the strengths and weaknesses of MC and CR items, (b) review alternative forms of MC items, and (c) introduce the explanation multiple-choice (EMC) items and discuss their unique features.

REVIEW OF MC AND CR ITEMS

Multiple-choice items are commonly used on large-scale standardized tests. The origin of MC items dates back to early in the 20th century when Frederick J. Kelly first introduced them in 1914 (Madaus & O'Dwyer, 1999). After that, MC items began to gain popularity because of their objectivity compared to essay questions. The first all-MC, large-scale test was developed to recruit military personnel in World War I and was known as the Army Alpha test (Madaus & O'Dwyer, 1999). The use of MC items was further promoted with the invention of the high-speed optical scanner in mid-1950s (Baker, 1971).

MC items have many advantages. Although it takes time and training to develop well-structured MC items, they can be easily administered and scored, thus becoming an effective way of measuring student knowledge on a large scale (Roediger & Marsh, 2005). MC items are also considered an objective form of assessment with high reliability (Wilson & Wang, 1995).

Limitations also exist for MC items. Because of their objectivity, MC items do not provide students with the opportunity to explain their answers, thus potentially limiting the depth and scope of information gathered from students. In science assessment, MC items tend to focus on discrete pieces of facts and have difficulties measuring certain aspects of inquiry science such as complex arguments or coherent understanding. MC items also fall short in eliciting student reasoning to explain or justify their choices. This lack of nuanced information about student reasoning may not be a concern for summative assessment, but for most classroom-based assessment it is important for teachers to understand student reasoning for instructional purposes. Wide application of MC items in classroom assessment may motivate teachers to emphasize superficial memorization of science facts rather than promoting deep scientific understanding (Linn & Hsi, 2000; Nichols & Sugrue, 1999; Resnick & Zurawsky, 2007).

CR items differ from MC items on both required student behavior and scoring objectivity (Rodriguez, 2003). Compared to MC items, CR items have the advantages of being able to provide a more direct assessment of what students know and can do on their own terms. CR items are also considered more authentic, as they provide opportunities for students to demonstrate a full range of abilities. In measuring student complex science understanding, CR items create a context for students to identify science ideas, provide their explanations of the science phenomena, and allow students to elaborate on their justifications using scientific evidence. Through analyzing student responses to CR items, teachers can identify student misconceptions and incoherent understanding to improve instruction (Black & Wiliam, 1998).

As much as CR items are welcomed by many science education researchers, they have their own constraints. They require more time to answer and cost more to score (Kennedy & Walstad, 1997; Livingston, 2009). Moreover, due to the involvement of human raters, CR items usually have problems with interrater reliability. CR items also tend to have lower score reliability, as the time required to complete CR items restricts the number of items that can

be contained in a single test. Research has shown that the correlation between the MC and CR items on a test is higher than the internal consistency of the CR items due to the high reliability of the MC items (Lukhele, Thissen, & Wainer, 1994). CR items are also much more costly than MC items. For example, on the Advanced Placement Chemistry test, it costs about $3 to $4 to score each CR item, whereas it costs less than 1 cent to score the entire set of MC items. To achieve the same acceptable reliability of .92, the scoring of the entire set of CR items (about 10) on each exam costs about $30 more than the scoring of the MC items for each test taker (Wainer & Thissen, 1993).

Although MC and CR items differ significantly in how they elicit responses from students (i.e., selection vs. generation), there is evidence of construct equivalence for these two item types. Through a large-scale review of studies investigating construct equivalence between MC and CR items, Rodriguez (2003) found that when the MC and CR items share the same item stem, their mean corrected correlation could be as high as .95. As Thissen, Wainer, and Wang (1994) pointed out, "recognition is not the same as generation, but they may be highly correlated" (p. 115). Research comparing both formats as measures of general cognitive constructs in standardized tests also confirmed the similarity between the two (Bennett, Rock, & Wang, 1991; Bridgeman & Rock, 1993; Klein et al., 2009).

ALTERNATIVE FORMATS OF MC ITEMS

Considering the differences and similarities between CR and MC items, researchers have explored alternative forms of MC items to improve their diagnostic function (Briggs et al., 2006; Sadler, 1998; Treagust, 1995, 2006). A common characteristic of these alternative items is that they ask students to provide justifications to their MC answers. For example, Treagust (1995) constructed two-tier MC items to measure student understanding of science concepts. Students first responded to a content question with two to three choices. They then selected from among four possible reasons explaining their answer to the first-tier question. The four reasons included explanations for the correct answer as well as incorrect answers. Results showed that a high percentage of students held alternative views of science topics that were different from those of teachers and scientists. These diagnostic instruments help teachers achieve better understanding of the nature of students' knowledge structure.

It has been common practice that sound test developers develop MC distractors based on student misconceptions. Sadler (1998) included distractors that represent common alternative science conceptions in MC items to measure student understanding of astronomy concepts. The purpose was to gather qualitative information on common student misconceptions without conducting large-scale, one-on-one interviews. Sadler used MC items with a stem and five choices. Only one of the choices was correct, and the rest were alternative conceptions. The alternatives were developed through either literature search or student interviews. Sadler found that students do not progress quickly from no knowledge to valid understanding. Instead, they may take small steps in reaching a coherent understanding. Sadler argued that distractor-driven MC items provide rich qualitative information, which helps teachers to diagnose student alternative conceptions and to help students move toward more integrated understanding.

The Briggs et al. (2006) study proposed an item format called Ordered Multiple Choice in which the MC categories are designed to reflect distinct levels of understanding of the

construct being measured. Ordered Multiple Choice items adopt a construct-driven approach by specifying student developmental stages on a construct. The Ordered Multiple Choice items are scored polytomously depending on the level of understanding the student achieved. The authors find that Ordered Multiple Choice items have great potential in providing diagnostic information at the classroom level with high reliabilities.

EXPLANATION OF MC ITEMS USED IN THIS STUDY

In this study, we continued the exploration of alternative MC items by designing EMC items to measure sixth and seventh graders' understanding of energy concepts such as energy source, transformation, and conservation. In our study, each MC item was followed by an EMC item to form a two-tier item. The MC items asked students to select from among four choices about a science phenomenon. The EMC items then asked students to select from among six choices to explain their answer to the previous MC item. The MC items were selected from published items from Trends in International Mathematics and Science Study (International Association for the Evaluation of Educational Achievement [IEA], 1995a, 1995b, 1999, 2003) on energy concepts. In previous research, we tested those MC items and asked students to explain their choice in a CR format (Lee, Varma, Linn, & Liu, 2010). In this study, we used student free responses to create choices for the EMC items. A detailed description of the EMC items is provided in the Item Design and Scoring section.

Although built on previous research, there are four major distinctions between the EMC items designed for this study and the other alternative MC items previously discussed. First, the distractors developed in Treagust (1995) and Sadler (1998) do not necessarily represent a progression of levels of understanding. Instead, those distractors may be parallel to each other and represent alterative views of science phenomena. The choices used in the EMC items in this study were designed to reflect distinct levels of understanding, ranging from more discrete, less connected explanations to more complex, more integrated explanations.

Second, most previous alternative MC items use four choices. Based on our analysis of student responses to previously administered CR items, the number of popular student views is often larger than four. Having only four choices may limit the possible explanations that students want to offer. In this study, we increased the number of choices in the EMC items from four to six. This allows for the inclusion of three choices targeting the correct first-tier answer and one choice targeting each of the incorrect first-tier answers. The design of six choices also reduces the chance of random guessing.

Third, the two parts in previous two-tier items have been scored together. Students receive the highest score only when they select the correct answer on both choices. Although this scoring method is the strictest way of rewarding students, it does not allow the examination of the relationship between the first- and second-tier answers, which is of key interest to us. In this study, we scored the two tiers separately to evaluate the consistency between first- and second-tier answers. The scores can be easily recoded and combined for other purposes if needed.

Last, although previous research (Briggs et al., 2006; Sadler 1998; Tamir, 1989; Treagust, 1989, 1995) demonstrated the value of alternative MC items, there is no empirical evidence directly comparing the information gathered from alternative MC items and traditional CR items. It is unclear whether the diagnostic information offered by the alternative MC items is

the same as the information gathered from student-generated responses. This study attempts to address this question through a random assignment. After answering each MC item, students within a class were randomly assigned to either an EMC item or a regular CR item based on the same item stem.

In this study, we compared the correlation between the MC items and EMC items and between MC items and CR items. We also compared the item difficulty of the MC, EMC, and CR items as in the two-tier format. We further examined whether students who were exposed to the EMC items have advantages in answering the MC items, as students were allowed to go back and change their MC answers. Finally, we investigated the alignment between item formats by examining the percentage of correct (or incorrect) answers on a CR item or EMC item given a correct (or incorrect) MC answer.

ITEM DESIGN AND SCORING

The development of the items used in this study were guided by the science knowledge integration (KI) framework (Linn & Eylon, in press; Linn & Hsi, 2000; Linn et al., 2006; Sisk-Hilton, 2009). KI represents a constructivist view of how science knowledge is acquired and refined. It is a view of cognition that emphasizes the multiple, diverse, and often contradictory ideas held by students about scientific phenomena (Linn, 1995; Linn, Davis, & Bell, 2004; Linn & Hsi, 2000). From the KI perspective, learning occurs when students take advantage of their own ideas, add new normative ideas, use scientific criteria to distinguish between the ideas, and form more coherent views of scientific phenomena. KI is based on the observation that one of the most important aspects of science is its generative capacity, the ability to solve problems by applying general concepts and principles. To advance knowledge, a scientist often has to elicit and link two or more appropriate concepts to solve a problem in a new situation. The KI framework emphasizes a repertoire of ideas that students build and refine while they interact with the real world in everyday settings and during science instruction. The KI framework takes advantage of the reasoning that students use to elicit, add, compare, and revise their ideas related to scientific phenomena. Using student misconceptions as the starting point, the KI framework describes science learning processes such as adding new ideas, distinguishing between new and existing ideas, developing scientific criteria to reconcile ideas, and building coherent links among relevant and normative ideas.

In this study, the KI framework guided both the development and scoring of the EMC items. As described earlier, 10 MC items were selected from the published Trends in International Mathematics and Science Study released item sets in 1995, 1999, and 2003 (IEA, 1995a, 1995b, 1999, 2003). These items address science content commonly taught in middle school such as the water cycle, food web, and chemical element recycling. These 10 MC items provided item contexts where students explained their choices. To each of the MC items we added an explanation part that asked students either to explain their choice (CR) or to select from among a list of provided explanations (EMC). Each EMC item has six choices, three providing explanations for the correct MC answer with progressing KI levels and one targeting each of the three incorrect MC choices. The six EMC choices were created based on a careful analysis of about 3,500 student responses to previously administered CR items (Lee & Liu, 2010).

In previous research, all CR items were scored using a rubric developed from the KI framework. The rubric has five levels (Liu, Lee, Hofstetter, & Linn, 2008):

- Irrelevant (score 0): Students' explanations did not include ideas that were relevant to the item context.
- No-link (score 1): Students' explanations were based on non-normative and scientifically invalid ideas.
- Partial-link (score 2): Students used scientifically normative and relevant ideas to the item context but did not elaborate how the two ideas were linked.
- Full-link (score 3): Students made a scientifically elaborated link between two normative and relevant ideas related to the item context.
- Complex-link (score 4): Students made two or more scientifically elaborated links between three or more normative and relevant ideas related to the item context.

Because a key interest of this study was to compare the MC/EMC tiers to the MC/CR tiers, we paid close attention to the alignment between the EMC items and the CR items in both the EMC choices and the scoring rubric. The three EMC choices targeting the correct MC answer were designed to represent the no-link, partial-link, and full-link KI identified in prior research. The reason that we did not create a complex-link level choice in the EMC items is that such choices are often considerably longer than other choices and may appear obvious to students as the right answer. As a result, the scoring rubric for the EMC items is 2 for the full-link choice, 1 for partial-link choices, and 0 for irrelevant and no-link answers. To ensure the comparability of the EMC and CR items, the original five-level scoring rubric for the CR items was modified to have the same three levels as the EMC items. See Figure 1 for a sample item set and scoring rubrics.

Two test forms were created for the comparison of the three item formats. Both forms contain the same 10 MC items. The second part of the item pairs, which is either a CR item or an EMC item, alternates between the two forms. For example, if an MC item is followed by an EMC item in one form, then the same MC item will be followed by a CR item in the other form. In each of the two forms, there are 10 MC items: 5 CR items and 5 EMC items. The two forms were administered online and were *randomly* assigned to students within a teacher.

PARTICIPANTS

The participants consisted of 794 sixth- and seventh-grade students taught by five middle school teachers in California. This study was part of a large research grant funded by the National Science Foundation in which the teachers were recruited to teach middle school energy topics. In this study, teachers volunteered to administer the assessment to all of their students. There were 343 (43.2%) sixth graders and 451 (56.8%) seventh graders in the sample, including 48.5% male students, 17.7% English language learners, and 70.1% with a home computer. The test was administered online at the end of a school year and took about 20 to 30 min to finish. Students within a teacher were randomly assigned to one of the two test forms. As a result, 52.3% of the students took Form A and 47.7% took Form B.

Tier 1 in Both Form A and B: The multiple-choice (MC) item

What is predicted to be a result of global warming?

(a) Rising ocean level
(b) More severe earthquakes
(c) Larger volcanic eruptions
(d) Thinning ozone layer

Tier 2 in Form A: The constructed-response (CR) item

Explain your choice. _____

Modified knowledge integration scoring rubric

Knowledge integration levels	Score	Description	Example Responses
Non-normative ideas	0	Restatement of the MC choice. Scientifically non-normative ideas or links	• I picked thinning ozone layer, because the amount of gases in the air are making our atmosphere thinner and everything on earth change, like the weather.
Normative ideas	1	One of the following ideas were present: • Temperature rising • Ice melting • Global warming cause	• The polar ice caps will melt and the oceans will receive much more water.
Linked ideas	2	Any number of links among the three ideas: • Temperature rising • Ice melting • Global warming cause	• As the temperature rises around the world, the ice caps will melt. The extra water will flow into the ocean and increase the water level greatly. • As fossil fuels are released in the atmosphere the heat gets trapped in the atmosphere, heating up the earth. This melts glaciers around the world and adds to the ocean water.

FIGURE 1 Scoring rubrics for the sample item set. (*continued*)

ANALYSIS

We applied a Rasch partial credit model (PCM; Masters, 1982) to analyze the assessment data. The choice of a PCM instead of a two-parameter model was made for three reasons: simplicity in test equating, availability of a discrimination index, and effective communication with teachers. Because two forms were used in this study, test equating was required to ensure the comparability of student performance. It is more straightforward to equate the forms based

Tier 2 in Form B: The explanation multiple-choice (EMC) item
Which of the following explains your choice?

(a) Global warming changes the intensity of ocean currents.
(b) More glaciers and ice caps melt as global temperatures get higher.
(c) Ocean receives water from melting iceberg.
(d) The Earth is getting hotter making tectonic plates move faster.
(e) Volcanic eruptions create a lot of heat.
(f) The ozone layer traps heat from escaping the Earth.

EMC scoring rubric

Knowledge integration levels	Score	EMC choice descriptions
Non-normative ideas	0	EMC-Choice (a) associated with MC-Choice (a) EMC-Choice (d) associated with MC-Choice (b) EMC-Choice (e) associated with MC-Choice (c) EMC-Choice (f) associated with MC-Choice (d)
Normative ideas	1	EMC-Choice (c) associated with MC-Choice (a)
Linked ideas	2	EMC-Choice (b) associated with MC-Choice (a)

FIGURE 1 (*Continued*).

on the difficulty parameter than on both the difficulty and discrimination parameters. Although the discrimination parameter is not included in the PCM, the software ConQuest (Wu, Adams, Wilson, & Haldane, 2007) used in this study provides a discrimination index for each item. It is the ratio of the difference between the average scores of the top 27% and bottom 27% groups divided by the maximum score allowed on the item. Values larger than .40 suggest good discriminating power and values less than .20 suggest poor discrimination (Ebel, 1954). This discrimination index is able to approximate the discrimination parameter from a two-parameter model (Kelley, Ebel, & Linacre, 2002). Finally, because the raw score is a sufficient statistic of the ability estimate in the Rasch PCM, teachers are more likely to understand the results in the form of ability estimates from the Rasch PCM than from a more complex two-parameter model.

We examined two important assumptions of the Rasch PCM: unidimensionality of the data and local independence of the items. In this study, the assessment was designed to measure a unidimensional science KI construct. Although the assessment items were embedded in different energy content topics, the KI scoring rubrics used in this study ensure that integrated science understanding is rewarded.

We used an exploratory factor analysis (EFA) to evaluate the dimensionality of the assessment. Specifically, we ran EFA using principal axis factoring with promax rotation. In addition, we used a technique called parallel analysis to determine the number of factors from the EFA. The underlying rationale for parallel analysis is that the eigenvalues of the salient factors extracted from the real data in an EFA should be larger than the eigenvalues of the corresponding factors generated from simulated random data (Horn, 1965). To conduct parallel analysis, we simulated a large number of data sets ($n = 500$) with the same sample size and the same number of variables as in our real data. We then compared the mean eigenvalues from the simulated data to the eigenvalues from the real data. The eigenvalues of the real data are expected to be larger than those of the simulated data, as meaningful and substantial factors should account for more variance than expected by chance. Parallel analysis has been well documented to be an effective way of detecting number of factors for the past 30 years (Carraher & Buckley, 1995; Horn, 1965).

The local independence assumption requires that the response to an item on a test be independent of the response to any other items after the level of attainment on the underlying construct is controlled for. In this article, the underlying construct of interest is science KI ability. If the local independence assumption is met, then the mean correlation between items, especially the ones sharing the same stem (e.g., MC and CR, and MC and EMC) should be close to zero after the KI ability is controlled for (Ferrara, Huynh, & Michaels, 1999; Ferrara, Michaels, & Huynh, 1995). To examine this assumption, we followed the Ferrara et al. (1995) method and divided students into 10 ability groups based on their ability estimates. We calculated the mean correlation of items with the same item stem for the MC/CR and MC/EMC pairs. Mean correlations equal to or below .03 are considered low and above .11 are considered high. The theoretical underpinnings of the Ferrara et al. (1995) correlational method are very similar to Yen's (1993) Q3 statistic for detecting local item dependence. In addition to examining the unidimensional and local independence assumptions, we also evaluated the fit between the Rasch PCM and the observed data.

Although the two types of paired items ask about the same science topic, the difference in item format (i.e., MC/CR vs. MC/EMC) in the two test forms may affect the difficulty of the items and thus affect student performance. Therefore, the two test forms were equated to ensure the comparability of the student ability estimates obtained from each form. The 10 MC items were used as the common items between the two forms in equating. The mean/sigma equating method was used to equate the two forms so that the item difficulties of the common items could be on the same scale (Kolen & Brennan, 2004). On the basis of the linear function obtained from the common items, the item difficulty of items on Form A was transformed to be on the same scale as the item difficulty of items on Form B. Student ability estimates on Form A were also transformed to be on the same scale as ability estimates obtained on Form B.

We examined the correlations between the MC and CR, and MC and EMC items. We conducted a chi-square analysis to examine whether student performance on the MC items was influenced by the subsequent CR and EMC items. The purpose of this analysis was to see whether exposure to the EMC choices gave students an advantage when answering the MC items. Finally, we investigated the alignment between different item formats by examining the percentage of correct answers on a CR or EMC item given a correct answer to a MC item. The correlation between the paired items was calculated, and a mean correlation was provided for the MC/CR and MC/EMC comparisons across the two test forms. The mean item difficulty of

the three item formats was calculated after the item difficulties of the items on the two forms were equated. We also used an analysis of variance to determine if there was any statistical significance in the item difficulties among the MC, CR, and EMC items.

RESULTS

Descriptive Statistics

Both Forms A and B showed reasonable internal consistency (Cronbach's $\alpha = .70$ for each form). The maximum score for both forms was 30. The mean score was 18.86 ($SD = 4.68$) for Form A and 19.25 ($SD = 4.93$) for Form B. There was no significant difference between the form scores ($p = .25$).

Dimensionality and Local Independence

The EFA results show that the first eigenvalue was 4 times as big as the second eigenvalue, and the second eigenvalue was not distinguishable in size from the rest of the eigenvalues. The first factor accounts for 58% of the variance in student scores. Results from parallel analysis confirmed the one-factor structure of the data (Figure 2).

Table 1 summarizes the results of the local independence examination. Most of the mean correlations were close to zero in their absolute values across the 10 ability groups. Seven of the 10 mean correlations were equal to or smaller than .03 for both the MC–CR and MC–EMC pairs. None of the mean correlations exceeded the .11 cut point (Ferrara et al., 1995). This finding provides evidence that the local independence assumption was met for the items used in this study.

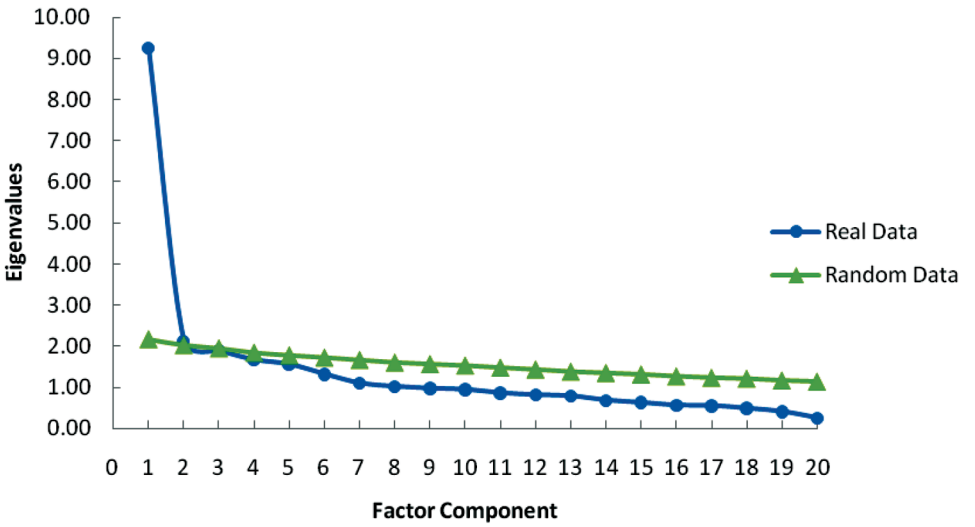

FIGURE 2 Eigenvalues from real data and from parallel analysis. *Note:* (Color figure available online).

TABLE 1
Results of Local Item Dependence

Ability Group	Theta Range		n	Mean Within-Tier Correlation	
	Low	High		MC–CR[a]	MC–EMC[a]
1	−1.77	−1.40	20	−0.02	0.04
2	−1.39	−1.02	33	0.03	0.02
3	−1.01	−0.64	52	−0.07	0.04
4	−0.63	−0.26	89	0.01	0.01
5	−0.25	0.12	126	0.03	−0.02
6	0.13	0.50	135	0.02	−0.06
7	0.51	0.88	129	−0.05	0.02
8	0.89	1.26	98	0.04	−0.03
9	1.27	1.64	62	0.03	0.01
10	1.65	2.02	50	0.03	0.02

Note. The mean within-tier correlation for multiple-choice (MC) and constructed response (CR) items is calculated based on 10 correlations between MC and CR items across the two test forms, and so is the mean correlation for the MC and explanation multiple-choice (EMC) items.

[a] $n = 10$.

Item Fit

The outfit statistic produced by ConQuest is used to evaluate the fit between the Rasch PCM and the observed data on each item. The outfit statistic detects unexpected student responses that are far below or above their ability estimates and has an acceptable range of .70 to 1.30 (Wright & Linacre, 1994; Wu et al., 2007). A small outfit value suggests that the item does not contribute to the measurement of the underlying ability beyond what is already measured by the rest of the items. A large outfit value suggests that the item fails to differentiate among students in terms of the target ability and thus may measure a different construct from the rest of the items. Obviously a large outfit statistic is more problematic than a small outfit statistic. The outfit statistics for the 20 items in this study fall between .89 and 1.14, with mean .99 and standard deviation .07. The finding supports the fit between the Rasch PCM and the empirical data.

Discrimination Index

Using the top and bottom 27% method described in the Analysis section, we obtained a discrimination index for each item. All of the items had reasonable discrimination values, with the lowest being .32 for an MC item. The mean discrimination value was .50 with a standard deviation of .18.

Student Ability and Item Difficulty Distribution

The Wright map shown in Figure 3 presents the distribution of student ability estimates on Form A and Form B, and the distribution of item difficulty by the three item formats. The x's

Logit	Ability Distribution	Item Distribution			Ability Distribution
	Form A	CR	EMC	MC	Form B
2	X	2.cr* 9.cr			
	Xxx				x
	Xxxx				xx
	Xxxx	1.cr			xx
	Xxxx				xxx
	Xxxxx		9.emc		xxx
1	xxxxxx	7.cr			xxxxx
		5.cr			
	xxxxxx	4.cr	4.emc	2.mc	xxxxx
	Xxxxxxxxxxx	3.cr		5.mc	xxxxxxx
		8.cr			
	Xxxxxxxxxxxxx	6.cr			xxxxxxx
	Xxxxxxxxxxx	10.cr	2.emc	7.mc	xxxxxxxxxx
			8.emc		
	Xxxxxxxxxxx		3.emc		xxxxxxxxxxx
0	Xxxxxxxxxx		5.emc	8.mc	xxxxxxxxxx
			1.emc		
			6.emc		
	xxxxxxxxxx		7.emc		xxxxxxxxx
	Xxxxxxxxx				xxxxxxxx
	Xxxx				xxxxxxxx
	Xxx			6.mc	xxxxxxx
				3.mc	
	Xxxx			4.mc	xxxx
-1	X		10.emc	9.mc	xxx
	x				xxx
	Xxx				xxx
	Xxx				x
	X				xx
	X				xx
-2					x
				1.mc	
				10.mc	

FIGURE 3 Items and student ability distribution. *Note.* CR = constructed response; EMC = explanation multiple-choice; MC = multiple-choice. (Color figure available online).

in the figure represent the students and the position of the x's indicates the ability estimate for that student. Each x represents 3.6 students for Form A and 3.5 students for Form B. The numbers in the third to fifth columns are the item difficulty estimates, organized according to the three item formats. For example, 2.cr represents the CR part of Item 2. The ability estimates

are calibrated to a logit scale, as are the item difficulty estimates. The higher the position, the more able the student is and the more difficult the item is.

The relative position between a student and an item determines the student's chance of getting that item correct. The further a student's ability estimate is above an item difficulty estimate, the more likely that the student will achieve the maximum score on that item. Similarly, the further a student's estimate is below an item estimate, the more likely that the student will fail to answer that item correctly. For instance, it will be extremely difficult for students whose ability estimate is at −2 on the logit scale (the lowest performing students) to answer correctly to the CR part of Item 2 (2.cr; the most difficult item on the test).

Both ability estimates and item difficulty estimates were equated between Forms A and B so they are comparable on the logit scale. The mean ability estimate of students who took Form A is .22 ($SD = .55$) and the mean ability estimate of students on Form B was .16 ($SD = .52$). Figure 3 shows no notable difference in the distribution of student performance between these two forms after equating. A t test of the difference between the mean estimate of students who took the two forms showed no statistical significance ($p = .08$), which was expected because the two forms were randomly assigned to students within a teacher.

Correlations Between the MC–CR Pairs and the MC–EMC Pairs

The second column in Table 2 shows the correlation between the paired MC and CR items. The highest correlation was .70 on Item 7 and the lowest was .13 on Item 1. The mean correlation between the 10 MC–CR pairs of items was .35. All of the correlation coefficients were statistically significant at the $p = .01$ level.

TABLE 2
Pearson Correlation Between MC and CR, and MC and EMC Items

	Pearson Correlation	
Item	*MC and CR*	*MC and EMC*
1	.13**	.30**
2	.14**	−.04
3	.56**	.50**
4	.25**	.47**
5	.60**	.17**
6	.49**	.42**
7	.70**	.63**
8	.16**	.37**
9	.25**	.32**
10	.22**	.51**
M	.35	.37
SD	.20	.18

Note. MC = multiple-choice; CR = constructed response; EMC = explanation multiple-choice.

**Correlation is significant at the .01 level.

The last column in Table 2 shows the correlation between the MC and the EMC items of the same pair. The highest correlation was .63 on Item 7 and the lowest was −.04 on Item 2. The mean correlation between the 10 MC–EMC pairs of items was .37. Nine of the 10 correlation coefficients were statistically significant at the $p = .01$ level.

Item Difficulty

The value of item difficulty ranges from −3 to 3. The higher the value, the more difficult the item is. The CR items are the most difficult in all cases (see the item estimates in Figure 3). We found a statistically significant difference in item difficulty (see Figure 4) among the three item formats through an analysis of variance, $F(2, 27) = 9.50$, $p = .001$. The eta-squared value (ratio of the sum of squares for item type to the total sum of squares) was .41, which means that item type contributed to 41% of the total variance, which is considered substantial. As expected, CR items were significantly more difficult than MC items ($p = .001$). There was no significant difference between EMC and the two other item formats ($p = .128$ with MC and $p = .103$ with CR). However, most of the EMC items were more difficult than the MC items of the same item stem (Figure 3).

Item 1 showed the largest difference in difficulty between the CR and EMC formats. The difficulty estimate was 1.52 for the CR format and −.19 for the EMC format. As the item characteristic curves indicate (Figures 4 and 5), as students' ability increased, the probability

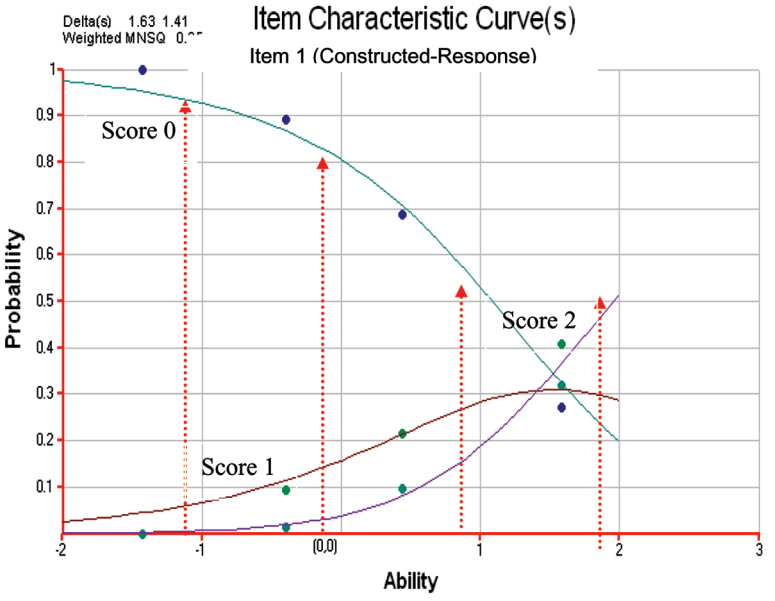

FIGURE 4 Item characteristic curve of the constructed-response version of Item 1. *Note.* The *x*-axis indicates student ability estimates on a logit scale from −2 to 2, and the *y*-axis indicates the probability of a student achieving a particular score given his or her ability estimate. The three curves in each of the two figures represent the three score levels (0, 1, and 2) used in this study, respectively. (Color figure available online).

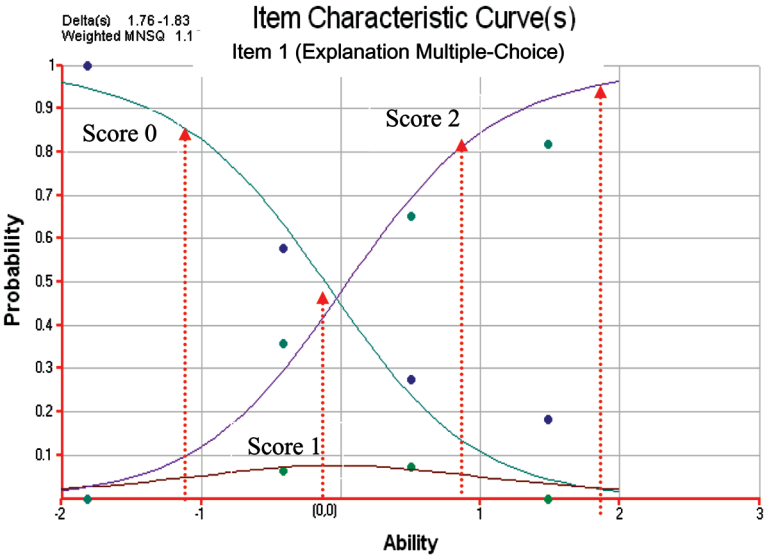

FIGURE 5 Item characteristic curve of the multiple-choice explanation version of Item 1. *Note*. The *x*-axis indicates student ability estimates on a logit scale from −2 to 2, and the *y*-axis indicates the probability of a student achieving a particular score given his or her ability estimate. The three curves in each of the two figures represent the three score levels (0, 1, and 2) used in this study, respectively. (Color figure available online).

of their achieving a lower score decreased and the probability of achieving a higher score increased. Students with a −2 ability estimate were likely to score 0 on both items as their probability of scoring 0 was .98 on the CR item (Figure 4) and .97 on the EMC item (Figure 5). At ability estimate −1, students had a probability of .92 of still scoring 0 on the CR item, but such probability reduced to .81 on the EMC item. As student ability increases, their probability of scoring 0 reduced to .46 at ability 0 and to .10 at ability 1 on the EMC item. However, for the CR item, the probability of scoring 0 was still very high at .80 at ability 0 and remained .50 at ability 1, which means that even the relatively proficient students were likely to provide an incorrect explanation to the CR item. Students with the highest ability at estimate 2 had a very high probability of .96 of scoring 2 on the EMC item, but only had a probability of .5 of scoring 2 on the CR item. The differences in the curves clearly explain the difference in the overall item difficulty between these two items.

Impact on MC Answers by Subsequent CR and EMC Items

We applied chi-square tests to examine whether student choice on MC items was affected by the subsequent CR or EMC item. The motivation for this analysis was a concern that the choices in an EMC item may provide a hint to students and help them answer the preceding MC items. If that was the case, students who responded to the MC–EMC item pairs would be expected to perform better on the MC part of the item pairs than students who responded to the MC–CR pairs. Table 3 shows no significant differences on 7 of the 10 MC items between the two

TABLE 3
Percentage of Students Who Chose Correct Multiple-Choice Answers When Paired With Generation and Selection Explanation Items

	CR		*EMC*			
Item	*N*	*Correct (%)*	*N*	*Correct (%)*	$\chi^2(1)$	*p*
1	415	92.05	378	91.29	.15	.70
2	416	36.86	377	39.32	.50	.48
3	378	69.66	415	72.29	.67	.41
4	415	72.77	376	64.64	6.11	<.05
5	377	18.46	413	38.07	37.19	<.001
6	376	73.87	410	69.88	1.56	.21
7	413	41.20	377	51.45	8.37	<.01
8	378	51.18	408	56.87	2.57	.11
9	376	73.88	409	73.25	0.04	.84
10	414	94.46	375	92.61	1.12	.29

Note. CR = constructed response; EMC = explanation multiple-choice.

kinds of item pairs. Three items did show statistically significant differences, though without a consistent pattern of favoring the MC–EMC students. Students who *selected* explanations outperformed students who *generated* explanations on the MC parts on Items 5 and 7. The opposite pattern occurred on Item 4. The results suggest that the choices in EMC items do not necessarily provide an advantage to students in solving MC items.

Alignment Between MC Answers and EMC Explanations

In this study, each EMC item had six choices. One choice related to each of the three incorrect answers on the preceding MC item. The rest of the three choices represented three KI levels explaining the correct answer on the preceding MC item. By alignment between MC answers and EMC explanations, we mean that students' choices on the EMC items should be relevant to their choices on the MC items. For example, if students choose an *incorrect* MC answer but choose an EMC explanation that targets the *correct* MC choice, then there is a misalignment between the MC and EMC choices. Table 4 shows the percentages of students whose MC answers were conceptually aligned with their EMC choices. The overall alignment percentages varied from 52.0 to 98.7% across items. A better alignment (80% or higher) was observed in six items (Items 1, 3, 4, 7, 9, and 10) where the same key words were present in both the MC and the EMC item parts. For example, Item 7 is about global warming. A choice for the global warming MC item, "Thinning ozone layer," is matched with a choice in the EMC item with "The ozone layer traps heat from escaping the Earth." When there was no apparent match in key words between the MC and EMC items, the overall alignment dropped to the 50 to 60% range as shown in the other items. Therefore, it is possible that students chose the corresponding EMC choices not because they understand the rationale but because they guessed on the key words. Future research is needed to clarify students' response behaviors on EMC items. For example, think-alouds can be conducted to elicit the reasons that underlie students' choices on the EMC items.

TABLE 4
Percentage of MC Answers Aligned With EMC Answers

	Correct MC Answer		*Incorrect MC Answer*		*Total*	
Item	*N*	*Aligned (%)*	*N*	*Aligned (%)*	*N*	*Aligned (%)*
1	345	96.2	28	64.3	373	92.6
2	149	72.5	228	38.6	377	52.0
3	300	95.0	115	73.0	415	88.9
4	245	97.1	131	86.3	376	93.4
5	158	86.7	255	45.0	413	61.0
6	288	76.7	122	36.1	410	64.6
7	194	97.9	183	71.6	377	85.1
8	236	84.7	172	22.7	408	58.6
9	303	90.8	106	50.9	409	80.4
10	351	100.0	24	79.2	375	98.7

Note. MC = multiple-choice; EMC = explanation multiple-choice.

An interesting pattern emerged when we separated the alignment analysis by the correctness of the answer to the preceding MC item. The pure chance of choosing a correct MC answer with an aligned EMC choice is about 12.5% (one out of four for choosing a correct MC answer times three out of six for choosing aligned EMC answer). The pure chance of choosing an incorrect MC answer with a matching EMC choice is also 12.5% (three out of four for choosing an incorrect MC answer times one out of six for choosing a matching EMC choice). Although all of the alignment rates were higher than the rate of pure chance, there was certainly variation across items, especially for students who chose an incorrect MC answer. Across all 10 MC items, the alignment between correct MC answers and matching EMC choices ranged from 72.5% (Item 2) to 100.0% (Item 10). In contrast, the alignment between incorrect MC answers and matching EMC choices ranged from 22.7% (Item 8) to 86.3% (Item 4). We conducted chi-square tests to examine the distributions of aligned and nonaligned EMC choices when correct and incorrect MC answers were chosen. All test results showed statistically significant differences in these distributions. Results showed that when students chose correct answers on the preceding MC items, they were more likely to select matching explanation choices than when they chose incorrect MC answers.

DISCUSSION

The EMC items presented here are early steps in learning to take full advantage of this alternative item format. The intention for designing EMC items is to elicit student justification and reasoning while preserving the objectivity of MC items.

Comparison of the item difficulty among EMC, MC, and CR items of the same item stem revealed that CR items are considerably more difficult than the EMC items of the same item stem. Figures 4 and 5 illustrate that on Item 1 even capable students cannot *generate* scientifically relevant explanations based on multiple ideas on their own but are able to *recognize*

such explanations when provided. When provided with explanations that are based on one single idea as compared to multiple ideas, they chose the multiple-idea explanation (score 2 in Figure 5) rather than the single-idea explanation (score 1 in Figure 5), making the single-idea explanation choice barely useful. On the other hand, when asked to generate explanations on their own, even high performing students were not able to generate multiple-idea explanations. Consistent with past research, students are more able to identify ideas than to generate ideas (Resnick & Resnick, 1992; Shepard et al., 2005). As a next step in our research, we will examine the item characteristic curves for all items to determine whether it is necessary to score EMC as 0, 1, 2 or whether a simple dichotomous score is sufficient as shown in Figure 5. Although significant, the mean correlations between the MC–EMC item pairs and MC–CR item pairs were not very high (.37 and .35, respectively). The EMC items were in general less difficult than the CR items, but they tended to be more difficult than corresponding MC items (Figure 4). The results suggest two findings: (a) When students select an answer on an MC item, it does not necessarily mean that they can explain the answer, which is suggested by the relatively low correlations between the MC items and the other two explanation item types, EMC and CR items, and (b) when explaining, it makes a difference whether the explanations are generated by students or provided to them as a choice. As shown in abundant previous research, *generation* is different from *selection* in science inquiry. Examination of the item characteristic curves of individual items (e.g., Figures 4 and 5) suggests that the response patterns are very different for students to reach different levels of scores on an EMC and a CR item. Note that the comparison of the difficulty of MC items in this study limits to MC items that were followed by CR or EMC items. The results may differ if stand-alone MC items were investigated.

One of the questions we had about the use of EMC items was whether they may provide a hint to students in their response to the corresponding MC items. This concern was not substantiated by the analysis results (Table 3). In general, the students who took the MC–EMC item pairs did not perform better than their peers who took the MC–CR pairs. This finding provides some validity evidence for the use of the EMC items when paired with preceding MC items.

Investigation of the alignment between MC choices and EMC choices showed a good alignment between correct MC answers and subsequent EMC answers. The degree of alignment varied across items for incorrect MC answers, with some items showing relatively poor alignment (Table 4). The better alignment for correct MC answers was probably due to the fact that there were three EMC choices targeting the correct MC answer but there was only one EMC choice for an incorrect MC answer. The alignment issue can be improved through a revised design of the MC–EMC item pairs. To elicit more useful information from students on their reasoning, we can take advantage of routed item design by creating an item branch for students choosing each of the four MC choices. In the new design (Figure 6), all the choices in a subsequent EMC item would be relevant to the choice students made in the preceding MC item. This way, the diagnostic value of the EMC items would be increased, as they now would have the potential to capture a range of reasoning students may offer. We would be able not only to detect the different reasons that students endorse the correct MC answer but also to identify misconceptions of varying understanding levels that underlie the incorrect MC answers.

A limitation of the current study lies in the limited number of item pairs (n = 10) used here. With the relatively small number of item pairs, the findings have limited generalizability. Another limitation lies in the common items based on which the test equating was conducted.

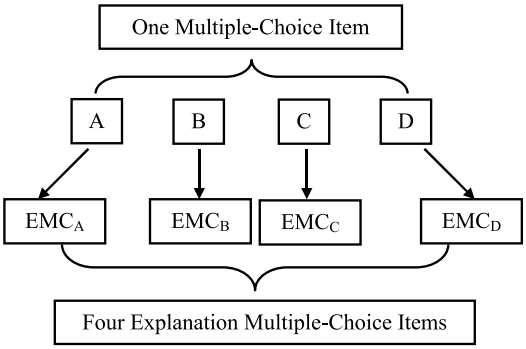

FIGURE 6 Flow chart of routed explanation multiple-choice items. *Note.* EMC = explanation multiple-choice.

Ideally, the common items should be representative of the test forms with regard to both item content and item types. Given the design of this study, only the MC items were available as common items.

Future research on EMC items will focus on two aspects: (a) exploring routed item branches as a means to increase the diagnostic value of EMC items and ameliorate the alignment issue, and (b) administering the improved version of EMC items at a larger scale both item wise and student wise. When more item pairs are tested, we are able to get a more complete picture of how EMC items function in measuring scientific reasoning. Currently we are not able to compute the correlation between the EMC and CR items because they are not administered to the same students. When students from more classes take the EMC items, we will be able to obtain the correlation between the EMC and CR items at the aggregated class level, which will provide direct evidence for the comparability of the EMC and CR items.

Findings from this study provide important implications for science educators and teachers. On one hand, this study provides direct evidence that EMC items function differently from regular CR items in measuring student science KI ability. Therefore, if the purpose of the assessment is to evaluate student ability in generating explanations for science phenomena, researchers and teachers should use CR items, as they provide the most direct measure of such ability. On the other hand, EMC items demonstrate great potential in eliciting diagnostic information as they provide a quick way for teachers to understand student alternative conceptions. For teachers who cannot afford the time to administer and score CR items, EMC items become an efficient way to probe popular student views about science topics. Based on the assessment development process of this study, we caution users of EMC items about the design of the EMC choices. The choices should reflect common student misconceptions for them to appear meaningful to most students. If such information is not readily available, teachers may want to gather popular student thoughts from CR items first and use the results to construct EMC items for later classes. When designing the choices, one should also consider the length of each choice to avoid making the correct answer too obvious to miss. Besides use for diagnostic purposes, we recommend that the MC–EMC pair items can also be used for summative evaluation purposes. Our analysis shows that the EMC choices do not necessarily help students answer the preceding MC item. The MC–EMC pair provides an efficient way to measure student understanding of science phenomena and their related explanations, and

therefore may be particularly useful in large-scale summative assessments for which efficiency is one of the primary concerns.

REFERENCES

American Association for the Advancement of Science. (1993). *Benchmarks for science literacy*. New York: Oxford University Press.

Baker, F. (1971). Automation of test scoring, reporting and analysis. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 203–234). Washington, DC: American Council on Education.

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement, 28*, 77–92.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*, 139–148.

Bridgeman, B., & Rock, D. (1993). Relationships among multiple-choice and open-ended analytical items. *Journal of Educational Measurement, 30*(4), 313–329.

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

Carraher, S. M., & Buckley, M. R. (1995). The effect of retention rule on the number of components retained: The case of cognitive complexity and the PSQ. *Proceedings of the Southern Management Association*, 474–476.

Clark, D., & Linn, M. C. (2003). Designing for knowledge integration: The impact of instructional time. *Journal of the Learning Sciences*, *12*, 451–493.

Ebel, R. L. (1954). Procedures for the analysis of classroom tests. *Educational and Psychological Measurement, 14,* 352–364.

Ferrara, S., Huynh, H., & Michaels, H. (1999). Contextual explanations of local dependence in item clusters in a large scale hands-on science performance assessment. *Journal of Educational Measurement, 36*, 119–140.

Ferrara, S., Michaels, H., & Huynh, H. (1995). *A beginning validation of causes of local item dependence in a large scale hands-on science performance assessment*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Heubert, J. P., & Hauser, P. M. (1999, April). *High-stakes testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179–185.

International Association for the Evaluation of Educational Achievement. (1995a). *TIMSS science items: Released set for population 1 (third and fourth grades)*. Chestnut Hill, MA: Boston College.

International Association for the Evaluation of Educational Achievement. (1995b). *TIMSS 1999 science items: Released set for population 2 (seventh and eighth grades)*. Chestnut Hill, MA: Boston College.

International Association for the Evaluation of Educational Achievement. (1999). *TIMSS 2003 science items: Released set eighth grade*. Chestnut Hill, MA: Boston College.

International Association for the Evaluation of Educational Achievement. (2003). *TIMSS science items: Released set for eighth grade*. Chestnut Hill, MA: Boston College.

Kelley, T., Ebel, R., & Linacre, J. M. (2002). Item discrimination indices. *Rasch Measurement Transactions, 16*, 883–884.

Kennedy, P., & Walstad, W. B. (1997). Combining multiple-choice and constructed-response test scores: An economist's view. *Applied Measurement in Education, 10*, 359.

Klein, S., Liu, O. L., Sconing, J., Bolus, R., Bridgeman, B., Kugelmass, et al. (2009, September 29). *Test Validity Study (TVS) report: Supported by the Fund for Improvement of Postsecondary Education (FIPSE).* Retrieved from http://www.voluntarysystem.org/index.cfm?page=research

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

Lee, H. S., & Liu, O. L. (2010). Assessing learning progression of energy concepts across middle school grades: The knowledge integration perspective. *Science Education, 94*, 665–688.

Lee, H. S., Varma, K., Linn, M. C., & Liu, O. L. (2010). Impact of visualization-based inquiry science experience on classroom learning. *Journal of Research in Science Teaching, 47*, 71–90.

Linn, M. C. (1995). Designing computer learning environments for engineering and computer science: The Scaffolded Knowledge Integration framework. *Journal of Science Education and Technology, 4*, 103–126.

Linn, M. C., Davis, E. A., & Bell, P. (Eds.). (2004). *Internet environments for science education*. Mahwah, NJ: Erlbaum.
Linn, M. C., & Eylon B.-S. (in press). *Science learning and instruction: Taking advantage of technology to promote knowledge integration*. New York: Routledge.
Linn, M. C., & Hsi, S. (2000). *Computers, teachers, peers: Science learning partners.* Mahwah, NJ: Erlbaum.
Linn, M. C., Lee, H.-S., Tinker, R., Husic, F., & Chiu, J. L. (2006). Teaching and assessing knowledge integration in science. *Science, 313,* 1049–1050.
Liu, O. L., Lee, H. S., Hofstetter, C., & Linn, M. C. (2008). Assessing knowledge integration in science: Construct, measures and evidence. *Educational Assessment, 13*, 33–55.
Livingston, S. (2009). *Constructed-response test questions: Why we use them; How we score them* (ETS R&D Connections, RD-11-09). Princeton, NJ: Educational Testing Service.
Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement, 31*, 234–250.
Madaus, G. F., & O'Dwyer, L. M. (1999, May). A short history of performance assessment. *Phi Delta Kappan,* pp. 688–695.
Masters, G. (1982). A Rasch model for partial crediting scoring. *Psychometrika, 49*, 359–381.
National Research Council. (1996). *National science education standards*. Washington, DC: Author.
Nichols, P., & Sugrue, B. (1999). The lack of fidelity between cognitively complex constructs and conventional test development practice. *Educational Measurement: Issues and Practice, 18*, 18–29.
Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educationa reform. In B. R. Gifford & M. C. O'Conner (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 37–76). Boston: Kluwer Academic.
Resnick, L. B., & Zurawsky, C. (2007). Science education that makes sense. *American Educational Research Association Research Points, 5*, 1.
Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement, 40*, 163–184.
Roediger, H. L., III, & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1155–1159.
Sadler, P. M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching, 35*, 265–296.
Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*, 4–14.
Shepard, L. A., Flexer, R. J., Hiebert, E. H., Marion, S. F., Mayfield, V., & Weston, T. J. (2005). Effects of introducing classroom performance assessments on student learning. *Educational Measurement: Issues and Practice, 15*, 7–18.
Sisk-Hilton, S. (2009). *Teaching and learning in public: Professional development through shared inquiry*. New York: Teachers College Press.
Tamir, P. (1989). Some issues related to the use of justifications to multiple-choice answers. *Journal of Biological Education, 23*, 285–292.
Thissen, D., Wainer, H., & Wang, X. B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement, 31*, 113–123.
Treagust, D. F. (1989). Development and use of diagnostic tests to evaluate students' misconceptions in science. *International Journal of Science Education, 19*, 159–169.
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). Mahwah, NJ: Erlbaum.
Treagust, D. F. (2006, September). *Diagnostic assessment in science as a means to improving teaching, learning and retention.* Paper presented at the 2006 National UniServe Conference (Symposium of Science Teaching and Learning Research). Sydney, Australia.
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*, 103–118.
Wilson, M., & Wang, W. C. (1995). Complex composites: Issues that arise in combining different modes of assessment. *Applied Psychological Measurement, 19*, 51–71.
Wright, B., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*, 370.
Wu, M., Adams, R. J., Wilson, M., & Haldane, S. (2007). ACER ConQuest 2.0 [Computer program]. Hawthorn, Australia: ACER.
Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187–213.