

Assessing Assessment: A Case of Teachers' Criteria for Quality Science Test Items

Ji-Eun Lee
Kyoung-Tae Kim
Oakland University

Abstract

This study investigated a case of ten in-service and pre-service teachers' criteria of assessing the quality of science test items. Utilizing 20 statewide standardized Earth and Space Sciences test released items, the participants were asked to evaluate the quality of test items by indicating numeric ratings and providing narrative justifications for their ratings. The purpose of this study was to probe what criteria the participants applied when reviewing the quality of test items without being informed of its source. Adapting Black and his colleagues' (1998, 2004) suggestion of the formative use of summative tests for students, this study utilized a summative test as a formative professional development opportunity by allowing participants to self-assess, reflect on, and question the quality of test items in a context that removed the external purposes of the test. The results showed that the participants of this study did not pay explicit attention to the content being tested. Rather, participants used a considerable variety of criteria, focusing on the presentation of the questions such as layout of the test, use of visual components, and word choices or clarity of directions, without referring to the concepts involved. The results offer insights into teachers' interpretations of quality science test items and issues for more targeted research.

Correspondence concerning this manuscript should be addressed to lee2345@oakland.edu.

Introduction

Teaching is a complex endeavor in which one has to consider many factors and people involved. Certainly, it is a complicated process to assess or measure the effectiveness of teaching and learning. Thus, it is no wonder why many controversial debates have subsequently taken place when it comes to the issue of assessment. The nation-wide accountability systems that rely heavily on standardized test scores in determining rewards, sanctions, or penalties have recently been the center of dispute. Although standardized testing is not a new trend in the field of education, the question about its impact on the quality of teaching and student learning has not clearly been answered.

In the overview of the literature on teachers' perceptions of state testing programs, Abrams, Pedulla, and Madaus (2003) summarized that statewide tests in high-stakes settings often oblige teachers to give greater attention to the tested content area, increase the pressure to improve student performance, decrease morale among teachers, and lead to a de-professionalization of educators. Black and Wiliam (1998), who advocated the need for formative assessment as an essential component of classroom

work, also noted the negative influence of short, external, summative tests on teaching practice. While admitting the important role these tests play in securing public confidence in the accountability of schools, Black and Wiliam (1998) warned that “such tests can dominate teachers’ work, and, insofar as they encourage drilling to produce right answers to short, out-of-context questions, they can lead teachers to act against their own better judgment about the best ways to develop the learning of their pupils” (p. 147). These results partially imply that many teachers are not given adequate opportunities to reflect upon how and why these tests (i.e., short, external, summative form of tests) appear as they are and the level of effectiveness these tests represent.

This study facilitated an opportunity for a group of in-service and pre-service teachers to review and reflect on the quality of sample state-wide standardized Earth and Space Sciences test items in a university graduate level course. This condition, a discussion in a university classroom, was deliberately planned in order to probe the participants’ views on the provided test items from a different perspective. Unlike the condition in their school settings where the participants had to be concerned about the results of students’ test scores and the subsequent issue of accountability, the setting in this study purely focused on the quality of the test items. To promote active engagement, the participants were not informed of the source of the test items at the time they reviewed the items. This setting allowed the participants to focus on their own thoughts and beliefs as they went through the review process to determine the quality of the test items.

It is important to note that the purpose of this study was not to measure the level of the participants’ Earth and Space Sciences content knowledge or change their beliefs about science teaching. Rather, it was for the opportunity to use the standardized test, which has been known for the pressure it inflicts on teachers, as a reflective tool for teaching. In the discussion of “assessment for learning in the classroom”, Black, Harrison, Lee, Marshall, and Wiliam (2004) suggested that through active involvement in the learning process, summative tests can be used in a formative way that helps students see themselves as beneficiaries, rather than victims, of testing because tests can help them improve their learning. This study applied the same logic to teacher professional development by focusing on the reflective review process and raising questions on how the test items might be improved. In addition, this study was designed to promote further discussion about teachers’ perceptions of quality science test question design and explore its ensuing implications.

Related Issues in the Literature

Criteria for Assessments: Ongoing Discussion

Assessment is a critical process in teaching and learning. There have been continued efforts to develop quality assessments and to search for more effective assessment methods. In determining the quality of any assessment criteria, reliability and validity are considered vital. In classical test theory, validity is defined as the “appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores” (American Educational Research Association, American Psychological

Association, & National Council on Measurement in Education, 1985, p.8). The reliability refers to consistency or repeatability of assessments. Although there is a consensus that the assessment accurately measures what is supposed to be measured in a consistent manner is considered a sound test, it is apparent that there is no one-size-fits-all type of assessment. The purpose of the assessment guides its format (e.g., summative assessment, formative assessment, alternative assessments). In this section, various criteria for quality assessments, especially in the area of science and mathematics, are reviewed without being limited to a specific form of assessment. The criteria for assessments reviewed in this section will be compared to the criteria proposed by the research participants of this study.

As a case for the criteria for comprehensive and summative assessment, the committee on indicators of precollege science and mathematics education (Murane & Raizen, 1988) reviewed test items according to two criteria: (1) importance – how important the knowledge being tapped is for a student, and (2) adequacy – how adequately the item tests particular knowledge, given the purpose of the test. The reviewers identified characteristics of high-quality science tests for national, state, or local assessment as follows (Murane & Raizen, 1988, pp.178-179):

- Assessment items should be based on a sampling of the ideal or desired curriculum in the subject area.
- Items should focus on central concepts for the course of grade level.
- The test should be designed from a matrix of desired learning objectives.
- A few items should offer new ways of thinking about a concept or solving a problem and provide topics for teachers to use in subsequent instruction.

Stokking, Van der Schaaf, Jaspers and Erkens (2004) investigated teachers' assessment practice at the classroom level for students' research skills in natural and social sciences. In their study, an expert panel evaluated teachers' assessment practices based on several quality criteria. Some of them are as follows:

- Are the teacher's assessment criteria comprehensible?
- Are the criteria formulated unambiguously?
- Are the criteria relevant in view of the learning goals the teacher is striving for with the assignment?
- Are the criteria suitable for assessing the correctness with respect to content?
- Are the criteria suitable for assessing the depth or quality level with respect to content?

Stokking et. al. (2004) claimed that there were grounds for concern about the clarity of teachers' assessment criteria, the consistency between teachers' goals, assignments, and the assessment criteria, and the validity of teachers' assessment practice.

In addition to the discussion on traditional tests, the body of research on new forms of assessment shows optimistic possibilities (e.g., Ashtiani & Babali, 2006; Rowe & Hill, 1996; Wolf, Bixby, Glenn & Gardner, 1991), even though the validity of the tests is still questionable (e.g., Stokking & Voeten, 2000). Regarding the merit of the alternative assessments, many studies claimed the possibility that these measure higher order skills better than traditional tests.

Teachers' Perceptions about the Effects of Statewide Testing

While there are positive and negative implications regarding statewide testing programs, a generally accepted viewpoint is that statewide testing and its associated policies greatly impact teachers' classroom practice and, ultimately, students' learning. A literature review by Abrams, Pedulla, and Madaus (2003) on teachers' perceptions about state-mandated testing programs raised the concerns about the perceived effect of high-stakes tests on the quality of education, as well as on teachers and students. Those concerns included: (1) State tests have a more powerful influence on teaching practice than do the content standards. Additionally, teachers perceive that state tests have led them to teach in ways that contradict their own notions of sound educational practice; (2) the substantial allocation of instructional time for specific test preparation by teaching test-taking skills and using test preparation materials and released items from the state; (3) the perceived human impact of the state test is an issue of concern, in particular the test-related pressure teachers and students experience. Based on their overview, Abrams, Pedulla, and Madaus (2003) suggested that it is imperative to expand the role of teachers in policy decision-making.

Other research studies also supported that teachers generally perceived the effects of standardized testing on students negatively. Several studies employing survey or interview methods reported that a large number of teachers surveyed or interviewed felt that high-stakes testing caused too much pressure and stress for students, and student morale had declined in response to the statewide high-stakes test (e.g., Clarke, Shore, Rhoades, Abrams, Miao, & Li, 2003; Jones & Egley, 2004; Jones, Jones, Hardin, Chapman, Yarbrough, & Davis, 1999; Koretz, Barron, Mitchell, & Stecher, 1996).

Pinder (2008) presented some mixed results in the report on Maryland's mathematics and science practitioners' perspectives on increased testing and the No Child Left Behind Act. While quantitative data findings suggested that there were more positive aspects of the statewide assessment, interview data indicated that teachers felt the test was ineffective, unrealistic, and did not meet the needs of students. Pinder (2008) suggested that more qualitative studies be conducted to explore the perspectives of the various states' practitioners, especially those in the fields of math and science.

Overall, the previous studies suggested that many teachers perceive the negative effects of high-stakes standardized testing over its positive aspects. Also, much discussion has been done regarding the teachers' perceptions about the external effects rather than teachers' view on the quality of the assessment itself.

The Formative Use of Summative Tests

Black and his colleagues (1998, 2004) claimed that formative assessment, as opposed to summative assessment like statewide standardized tests, is an essential component of classroom work and its development can raise standards of achievement. However, they acknowledged that all teachers are required to undertake some summative assessment mainly for external purposes. Black and his colleagues suggested achieving a more positive relationship between formative and summative assessment. As a way to build such a positive relationship, the formative use of summative tests was considered. Some possibilities included facilitating students' engagement in a reflective review of the work and plan their revision, encouraging students to set questions and mark answers to gain an understanding of the assessment process, and encouraging students' involvement in peer assessment and self-assessment to apply criteria to help them understand how their work might be improved. The key message in this line of research is that summative tests can be a positive part of students' learning process by actively engaging students in the testing process.

Situating the Study

The design of this study was informed by the results and implications from prior research. First, as shown in many studies, the discussion on the criteria for sound assessments is not a new arena of research and continues to evolve. This study probed what kinds of criteria the participants of this study apply when they review the quality of statewide Earth and Space Sciences test items without being informed of their source. This context helped the participants review the items based on their own knowledge and beliefs, not based on the assumption that standardized test items developed by experts must be superior. In addition, considering the fact that numerous research studies report on teachers' perceptions on the effects of high-stakes testing, this research context facilitated new discussion on the teachers' perceptions on the quality of actual test items.

Second, adapting Black and his colleagues' (1998, 2004) suggestion of the formative use of summative tests for students, this study utilized a summative test as a formative professional development opportunity by allowing participants to self-assess, reflect on, and question the quality of test items in the context that removed the external purposes of the test.

Method

Participants

Ten graduate students enrolled in an Earth Science content course within a graduate level integrated science program, in a small-sized, private, Midwestern

university participated in this study. Nine participants were in-service teachers at K-8 class settings and one participant had no teaching experience at the time of the study. Table 1 briefly shows each participant's background and experiences. The instructor for the course was the second author of this study.

Table 1
Participant Demographics

Participant Number	Gender	Teaching Experience	Teaching grade level
P1	Female	more than 30 years	1 – 4
P2	Female	10 years – 15 years	5
P3	Female	10 years – 15 years	4 – 5
P4	Male	10 years – 15 years	7
P5	Female	less 5 years	6 – 8 (math)
P6	Male	10 years – 15 years	5 – 8
P7	Female	none	n/a
P8	Female	10 years – 15 years	Kindergarten
P9	Female	10 years – 15 years	7 – 8
P10	Female	15 years – 20 years	Science

Context

As part of the class work, the participants engaged in an activity that consisted of two tasks: (1) Task 1: Individual ratings of the quality of provided science test items, and (2) Task 2: Partner interview on the justifications for individual ratings.

Task 1: Individual ratings. For the first task, the participants were given a questionnaire which consisted of 20 questions taken directly from the 2006-2007 fifth and eighth grade Ohio Achievement Test (OAT) released items in Earth and Space Sciences (see Appendix). Before starting their individual ratings, it was emphasized that what was asked for was the participants' personal judgment. The participants were asked to rate how each test item represented the high quality science test in which it belongs on a 5-point scale with five being the highest quality. After providing directions, the 20 OAT Earth and Space Sciences test items were provided. The instructor did not inform the participants that these items were taken from the OAT, thus intentionally minimizing the possible influence from knowing the authority of the source. For Task 1, the participants rated and solved the questions individually. Task 1 was completed in about 30 minutes.

Task 2: Partner interview. After completing Task 1, the participants were randomly paired up by the instructor. The instructor of the class frequently used small group activities or partner work in the course. Thus, the participants were familiar with this type of discussion.

The purpose of Task 2 was to share the participants' justifications for ratings in Task 1 with their partner interviewers. For example, when Participants A and B were paired up, Participant A first took the role of an interviewer and Participant B was an

interviewee. Participant A (interviewer) asked Participant B why he or she rated the given test item in specific number and Participant B (interviewee) provided his or her justification. When Participant B finished explaining, they switched roles, Participant B being an interviewer and Participant A being an interviewee. The interviewer's job focused on identifying the interviewee's criteria of good science problems. The interviewees were allowed to freely express their justifications addressing many criteria. However, they were required to address the most vital criterion for their rating at the end of each item review. Whether or not the answers for the problems were correct was not discussed in the interview session. After individual interviews, the interviewers wrote a summary of the interviewee's criteria of good science problems. It took about two hours for the participants to complete the interviews with their partners.

Data Source and Analysis

This study adopted an exploratory and descriptive nature, offering plausible explanations for further investigation of quality test items in science teacher education based on the case of the participants of this study (Yin, 1994, 2006). To display the evidence obtained from the case of this study, the participants' individual rating worksheets and partner written interview record sheets were collected. Individual rating worksheets provided numeric evaluations. The interview record sheets provided justifications behind the rating. The data were analyzed via multiple levels: (a) analysis of the quantitative data from Task 1, and (b) analysis of the qualitative data from Task 2.

The participants' individual ratings demonstrated their overall judgment on the quality of provided test items. However, we were aware that the participants' individual ratings represented their subjective judgment rather than absolute norm. Especially, we found that a rating of '3' was used to indicate both good and poor items for some cases. In these cases, the participants' justifications for their ratings were considered in the interpretation of the numeric ratings.

To analyze the qualitative data from Task 2, we followed a double-coding procedure. We individually reviewed the data collected and identified frequently employed criteria by the participants for the justification data. Because ten participants provided their justifications for 20 test items, each author independently analyzed a total of 200 written statements altogether. Authors then jointly synthesized their individual findings. The major themes in the participants' justifications included the following: (1) higher order thinking/critical thinking skills, (2) visual representation, (3) grade level standards, (4) students' prior knowledge, (5) connection to other subjects or real-life, (6) directions/word choices, (7) layout of questions/format of questions, and (8) essential science concepts.

Scope and Limitations of the Study

By utilizing the case study methodology, this study was descriptive and exploratory rather than explanatory. The scope of this case study was limited to exploring and describing the ten participants' cases in a graduate level Earth Science course setting. Thus, the following limitations should apply when interpreting the

research findings. First, the small sample size of the study limits the ability to generalize the findings to a larger population of teachers. Second, despite finding individual differences between participants, it is inappropriate to infer variances between different subgroups (e.g., based on gender, number of teaching experiences) from the results of this study due to the small sample size. Third, the discussion of this study is limited to what the participants' wrote and stated in a university classroom. It is not guaranteed that the participants' assessment criteria directly represent their assessment practice in the classroom. Finally, since this study used a statewide standardized test, which mainly consisted of multiple choice type questions, this study did not attempt to generalize the participants' views to other forms of assessment.

Results

Table 2 presents the data from Task 1 summarizing the participants' individual ratings for each test item. Question #3 showed the lowest mean score (2.35) and Questions #1, #8, #17 showed the highest mean score (4.30).

Table 2.
Participants' Individual Ratings

Question	Participant	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Mean
Q1		4	5	5	4	4	4	5	3	4	5	4.30
Q2		3	2.5	4	4	4	4	2	5	5	4	3.75
Q3		2	2.5	3	3	1	1	3	2	3	3	2.35
Q4		2	2.5	4	4	3	5	3	2	4	4	3.35
Q5		3	2	4	4	5	5	4	5	4	4	4.00
Q6		2	2	3	3	3	3	4	2	2	3	2.70
Q7		3	2	5	3	2	4	3	4	5	5	3.60
Q8		5	5	4	4	5	5	3	3	5	4	4.30
Q9		5	5	4	3	4	4	3	1	5	2	3.60
Q10		5	5	5	4	4	4	3	5	4	3	4.20
Q11		5	5	3	4	2	2	4	5	2	4	3.60
Q12		5	5	2	1	2	2	4	4	1	2	2.80
Q13		5	5	2	2	3	3	4	4	1	3	3.20
Q14		2	5	4	4	4	3	4	3	4	4	3.70
Q15		2	5	2	3	3	4	4	3	5	4	3.50
Q16		4	5	3	3	5	3	4	5	2	4	3.80
Q17		3	5	4	4	4	5	5	5	3	5	4.30
Q18		4	5	3	4	2	3	4	5	2	4	3.60
Q19		2	5	4	4	3	4	4	4	4	4	3.80
Q20		2	5	3	4	5	3	4	5	3	4	3.80

The data from Task 2 were analyzed based on the themes identified in the independent and joint analyses of a total of 200 written justifications by two researchers. Some of the participants' justifications addressed multiple categories. However, we focused on the most vital criterion the participants proposed in each response, as directed

before starting Task 2, and attempted to avoid multiple coding. Twelve responses (6%) provided only numeric scores without written justifications. Thirty-six justifications (18%) were not included in these categories due to one of the following reasons: (a) no specific reasons were provided (e.g., “not great but a good question,” “did not like the question”) or (b) explanations were too broad or vague (e.g., terms such as “typical question,” “supporting students,” and “awareness of science”). Thus, 152 justifications (76%) were categorized into one of the major themes.

Table 3 presents the major criteria in the participants’ justifications and whether those criteria were used in the justification of good or poor quality test items.

Table 3.
Major Themes in Participants’ Justifications

Major Themes/Criteria	Used to describe good quality test items [number of responses (%)]		Used to describe poor quality test items [number of responses (%)]	
Higher order thinking/critical thinking skills	25	(12.5%)	20	(10%)
Visual representation	17	(8.5%)	14	(7%)
Directions/word choices	2	(1%)	23	(11.5%)
Prior knowledge	16	(8%)	0	(0%)
Lay-out of questions/format of questions	1	(0.5%)	13	(6.5%)
Connection to other subjects or real-life	5	(2.5%)	5	(2.5%)
Essential scientific concepts	5	(2.5%)	2	(1%)
Grade level standards	4	(2%)	0	(0%)

More detailed descriptions of the noted categories along with examples of participants’ written responses are as follows.

Higher Order Thinking/Critical Thinking Skills

One of the most frequently used phrases to describe good quality items was “higher order” or “critical thinking” skills. Four participants extensively used these terms to assess the quality of test items. An excerpt is as follows: “...applied to some of Bloom’s Taxonomy.... that higher order thinking skills will be applicable” (Question #1, Participant #1). It was not clear whether there was a consensus on the meaning of higher-order thinking among the participants beyond the educational cliché. A notable example of this occurrence was when two participants perceived the quality of Question #7 differently using this criterion. Participant #7 negatively stated, “[this question] does not

require much thinking,” whereas Participant #9 positively evaluated, stating, “[this is] a good problem solving question.”

Visual Representations

About 8.5% of the justifications indicated that visual representations, such as charts, diagrams, and photographs were perceived by participants as a component of good quality science test items. Using the same reasoning, about 7% of responses used this criterion to describe poor quality test items; responses pointed out the absence of visual representation or the low quality of visual tools that were provided. The following are some examples:

“...gives a visual which allows [students] to make an analogy of future events” (Question #13, Participant #6)

“...need more of an illustration for a visual learner.” (Question #4, Participant #5)

Directions/Word Choices in the Directions

Twenty-five responses indicated that directions or word choices in the directions were one of the important criteria in participants’ perceptions of the quality of science test items. Except for two justifications, all responses in this category used these criteria to point out the low quality of the test item. Some responses broadly mentioned that there was too much or too little information in the directions, while others specifically suggest different word choices. Some examples include:

“...question didn’t contain enough information, too vague” (Question #6, Participant #8).

“...difficult to understand, vague choice, need to be more specific: less than $\frac{1}{4}$ ‘of’” (Question #3, Participant #10).

“...don’t like the word before, why not use ‘during’?” (Question #6, Participant #4).

“...needs to include the distance in the question” (Question #18, Participant #5).

Prior Knowledge

Approximately 8% of justifications used the presence of prior knowledge in the test item as a criterion of a good quality test item; however there was a subtle difference among the justifications. Some participants gave high scores because “students had to draw from knowledge developed/gained during class to solve the question” (Question #10, Participant #8). This justification strongly implies that the participant saw the question as an opportunity for students to synthesize their prior knowledge. Another example mentioned this prior knowledge as a “fact covered in class.” Participant #8

rated Question #5 at the highest score since “students would be required to recall learned knowledge of the unit in order to determine a correct answer.”

Layout of Questions/Format of Questions

Fourteen justifications were based on the layout of the test and format of the questions. Once again, except for one response, all responses in this category used these criteria to justify the low quality of the test items. Participants #3 and #4 rated Questions #12 and #13 on the low end of the scale because these questions were not on the same page with the map in the questionnaire (see Appendix A for the layout of the questionnaire used in this study). Participant #8 extensively considered the format of the questions as a critical justification tool, while his numeric scores were at a higher rating than 3. Participant #8 also suggested using different question formats for Questions #14-#15, #19-#20, indicating his dissatisfaction. Participant #8 suggested employing short essay questions instead of multiple choice questions for all of these test items, since “a written response would have shown acquired knowledge in a better way.”

Connection to Other Subjects and Real-Life

Ten responses considered the connection to other subjects, or real-life, in their justification citing both positive and negative aspects. Some positive justifications include the following: “mentions geography of states, which helps students understand the erosion process” (Question #10, Participant #5) and “This question requires reading skills” (Question #8, Participant #3). Most of the negative justifications in this category addressed related mathematics content. This strongly implied that the mathematics content used in the test could be an obstacle. Some excerpts showing this are as follows: “Many students learn that $\frac{3}{4}$ of the earth is covered by water and may not be able to convert it to percentage” (Question #4, Participant #3) and “Math skills of fractions may throw off understanding of answer” (Question #3, Participant #4).

Essential Scientific Concepts/Skills

Seven responses stated that questions were good or poor since they were asking about important scientific concepts or skills. Unlike the responses in the grade level standards category, these statements were rather general. Participant #2 noted, “It is important to know information regarding rock formation process” (Question #14).

Grade Level Standards

Although no information was given on the source of tests or grade levels, four responses stated that the questions were good because they fit the grade level they were teaching. For example, Participant #3 considered Question #7 very good because it “fits with the 4/5th grade standards very well.”

Other Aspects Noted

The above eight categories scan overall perceptions of this group of participants. Although it is beyond the scope of this study, we would like to address some aspects that

might not be as evident in this overview. First, each individual tended to have one or two dominant preferred criteria. For example, Participant #1's justifications were primarily based on higher-order thinking and visual representations, even though the meaning of higher-order thinking was used in a vague manner. The same was true for Participant #2's "essential scientific concepts," and Participant #5's "visual representations." Secondly, whether the arguments were correct or not, only Participant #2 provided two justifications based on the specific content the questions were asking. Two examples are as follows:

"...it's 50%-50% daytime and night time. However, night time changes throughout the seasons, so it should include what part of the year being referred to" (Question #3).

"...erosion doesn't only take place on ground – reiterate this" (Question #10).

Discussion and Implications

This study reported on the design, implementation, and analysis of a study that centered on the ten participants' perceptions of good science test items and included their justifications. Using a statewide standardized test as a tool in a blind review format, this study attempted to promote the opportunities for the participants to focus on their own beliefs about quality assessment. While some of the participants' criteria for good science test items overlapped with the characteristics of high-quality science tests identified by other studies as presented earlier, several aspects noted in this study appeared to bring some implication for further discussions.

First, in analyzing the numeric ratings (Task 1) and narrative justifications (Task 2), we found that participants of this study did not pay explicit attention to the content being tested; those that were considered important characteristics of quality science tests and mentioned in the previous studies (e.g., Murane & Raizen, 1988; Stokking, Van der Schaaf, Jaspers & Erkens 2004). Rather, participants showed a considerable variety of criteria focusing on the presentation of the questions. A total of 70 written justifications (46% of the narrative data analyzed) showed evidence that the participants assessed the quality of test items based on the overall layout of the test, use of visual components, and word choices or clarity of directions without referencing the concepts involved. The reasons for this situation were not clearly addressed in the scope of this study. However, as indicated in prior studies on pre-service and in-service teachers' subject matter knowledge, weak content preparation and lack of confidence in their subject area knowledge (Atwood & Atwood, 1996; Abd-El-Kjalick, Bell, & Lederman, 1998; Brickhouse, 1990; Lederman, Gess-Newsome, Lantz, 1994) might have prevented participants of this study from focusing on the content being assessed, subsequently causing them to just beat around the bush.

Second, previous studies support that statewide testing and its associated policies greatly impact teachers' classroom practice and student learning. Additionally, teachers' views on these testing programs are not quite positive, particularly in relation to the rewards or penalties that are possible as a result of the test (e.g., Abrams, Pedulla, &

Madaus, 2003; Pinder 2008). Regardless of the quality of the test items, teachers feel that they have to “teach to the test,” in part due to its consequences. This top-down approach has been a source of concern as educators call for expanding the teachers’ role within the policy decision-making process. By design, the participants of this study were not informed that the test items were taken from the state-mandated test to elicit participants’ own judgments. Participants’ numeric scores and written narratives varied between extremes. In many responses, participants perceived that the given test items should be improved to better serve their students and teachers. This result supports the previous studies indicating the lack of classroom teachers’ participation in the decision-making process and the shortcomings of the top-down approach. This leads us to a dilemma that while participants of this study felt that some of the given test items were ineffective, as addressed above, the major criteria they used are still a source of concern. If the participants’ weak subject matter knowledge is the main reason for creating a set of criteria other than the content assessed, the validity of their judgment would be in question.

Third, we found that a few key words appeared in the participants’ justifications, terms such as “critical thinking” or “higher order thinking,” as important criteria used to decide the quality of test items. Although these are important characteristics of quality test items, what is unclear within the scope of this study are participants’ own definitions and how they embody this component in their actual teaching practice. In other words, what we can say is one thing, while what we know or can do is often another. According to Banilower, Smith, Weiss, and Pasley (2006), many teachers say their instruction is aligned with National Science Education Standards, but few demonstrate this in their teaching. Likewise, participants frequently mentioned critical thinking or higher-order thinking, when it is not clear where there is a consensus defining this term beyond the level of cliché.

Although the three aspects addressed above left more questions to probe rather than definite answers, the participants shared one positive experience out of this study. When the data were analyzed and shared with the participants in an informal debriefing session, there was a general consensus among the participants. They agreed that they, as teachers, must put some conscious effort into developing and critiquing assessment items. In particular, they admitted that they seldom considered evaluating or critiquing the quality of external test items as their responsibility. The participants also noted that they perceived many of the given items were not high quality for different reasons unlike their previous assumptions. However, the participants also commented on their own evaluation criteria when they were asked how critical the criteria were. They felt that all the criteria they used together certainly could make good test items. However, they were less confident in discussing the hierarchy of the criteria (e.g., “How important is it to consider the layout of test? Is it more important than providing clear directions?” “Is it necessary to incorporate visual representations in science tests? Why? How viable are your justifications?”). Even though many questions remained unanswered, the participants’ opportunities to evaluate test items and to discuss the quality of items with their peers raised the participants’ awareness about the need to examine their assessment practice in order to be explicit about what they were doing. Black and his colleagues (2004) suggested building a more positive relationship between formative assessment and

summative assessments. They proposed the formative use of summative tests for students by promoting students' engagement in the reflective review process of peer- or self-assessments and encouraging students to set questions. As the summative tests can be a positive part of students' learning process through active engagement, the context of this study demonstrated that the analysis of the summative tests can be a positive part of teachers' professional development by actively communicating and discussing the assessment criteria, which ultimately helps the participants to clarify the goals for teaching and to get instructive feedback.

We believe that this study provides both the participants and the instructor with opportunities for mutual reflection. It is worth noting participants' interpretations of quality science test items that might be different from what many science teacher educators expect. The results of this study also lead to some issues and questions to further probe: What if the participants were aware of the authority of the source prior to assessing each item? What if participants were asked to write test items by themselves? What if the test items were for other subject areas? While this study illustrates a snapshot of a group of participants' views, follow-up studies of a larger, more diverse sample of teachers will help to probe answers for the more targeted questions addressed above.

References

- Abd-El-Khalick, F., Bell, R., & Lederman, N. (1998). The nature of science and instructional practice: Making the unnatural natural. *Science Education*, 82, 417-436.
- Abrams, L. M., Pedulla, J. J., & Madaus, G. F. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory into Practice*, 42(1), 18-29.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, D.C.: American Psychological Association.
- Ashtiani, N. S., & Babaii, E. (2006). Cooperative test construction: The last temptation of educational reform? *Studies in Educational Evaluation*, 33, 213-228.
- Atwood, R. K., & Atwood, V. A. (1996). Preservice elementary teachers' conceptions of the causes of seasons. *Journal of Research in Science Teaching*, 33, 553-563.
- Banilower, E., Smith, S., Weiss, I., & Pasley, J. (2006). The status of K-12 science teaching in the United States. In D. Sunal & E. Wright (Eds.), *The impact of state and national standards on K-12 science teaching* (pp. 83-122). Greenwich, CT: Information Age.
- Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80(2), 139-144, 146-148.
- Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan*, 86(1), 8-21.
- Brickhouse, N. W. (1990). Teacher's beliefs about the nature of science and their relationship to classroom practice. *Journal of Teacher Education*, 41, 53-62.
- Clarke, M., Shore, A., Rhoades, K., Abrams, L., Miao, J., & Li, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from interviews with educators in low-, medium-, and high-stakes states*. National Board on Educational Testing and Public Policy, January 2003, Chestnut Hill, MA. (ERIC Document Reproduction No. ED 474867).
- Jones, B.D., & Egley, R. J. (2004, August 9). Voices from the frontlines: Teachers' perceptions of high-stakes testing. *Education Policy Analysis Archives*, 12. Retrieved February 19, 2011, from <http://epaa.asu.edu/ojs/article/view/194>.

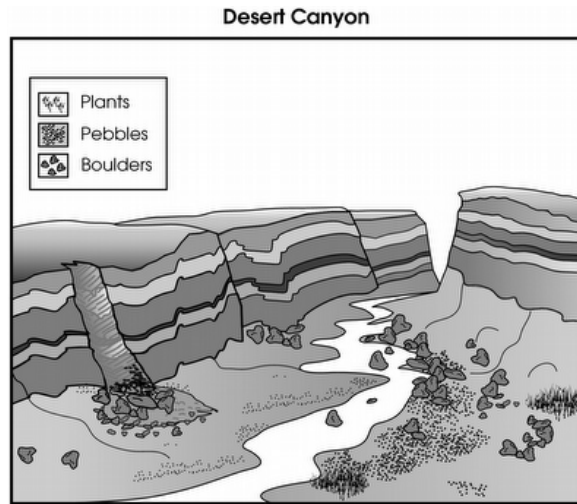
- Jones, M. G., Jones, B.D., Hardin, B., Chapman, L., Yarbrough, T., & Davis, M. (1999). The impact of high-stakes testing on teachers and students in North Carolina. *Phi Delta Kappan*, 81(3), 199-203.
- Koretz, D., Barron, S., Mitchell, K., & Stecher, B. (1996). *The perceived effects of the Kentucky instrumental results information system (KIRIS)*. Retrieved February 19, 2011, from http://www.rand.org/pubs/monograph_reports/2007/MR792.pdf.
- Lederman, N. G., Gess-Newsome, J., & Latz, M. S. (1994). The nature and development of preservice teachers' conceptions of subject matter pedagogy. *Journal of Research in Science Teaching*, 31, 129-146.
- Murane, R. J., & Raizen, S. A. (Eds.). (1988). *Improving indicators of the quality of science and mathematics education in grades K-12*. Washington, D.C.: National Academy Press.
- Pinder, P. J. (2008, January). *A critique analysis of NCLB, increase testing, and past Maryland mathematics and science HAS exams: What are Maryland practitioners' perspectives?* Paper presented at the International Conference of the Association for Science Teacher Education. St. Louis, MO. (ERIC Document Reproduction Service No. ED499781)
- Rowe, K. J., & Hill, P. W. (1996). Assessing, recording and reporting students' educational progress: The case for 'subject profiles.' *Assessment in Education*, 3(3), 309-352.
- Stokking, K., Van der Schaaf, M., Jaspers, J., & Erkens, G. (2004). Teachers' assessment of students' research skills. *British Educational Research Journal*, 30(1), 93-116.
- Stokking, K., & Voeten, R. (2000). Valid classroom assessment of complex skills. In R. Simons, J. Van der Linden & T. Duffy (Eds.), *New learning* (pp. 101-118). Boston, MA: Kluwer.
- Wolf, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment, *Review of Educational Research*, 17, 31-74.
- Yin, R. K. (1994). *Case study research: Design and methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Yin, R. K. (2006). Case study methods. In J. L. Green, G. Canolli & P.B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp.111-122). NJ: Mohwah Lawrence.

Appendix. Questionnaire for individual ratings

Category: GOOD SCIENCE TEST ITEMS

Please indicate your rating by circling an appropriate number. Provide your answer for the problem.

Poor ←-----→ Good
 Example Example
 (1) (2)

Question #1: Rating (1 2 3 4 5)

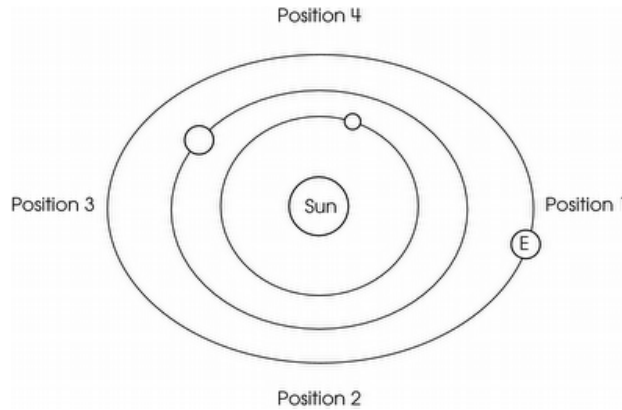
The picture shows a stream flowing through a desert canyon. The canyon was shaped by natural processes.

Identify a slow process that could have helped shape the canyon. Describe evidence of this process shown in the picture.

Then, identify a rapid process that could have helped shape the canyon. Describe evidence of this process shown in the picture.

Question #2: Rating (1 2 3 4 5)

The diagram shows the position of Earth (E) now.



Where will Earth be in six months?

- A. near position 1 B. near position 2
C. near position 3 D. near position 4

Question #3: Rating (1 2 3 4 5)

A class was learning about the pattern of day and night on Earth.

What part of Earth experiences night at the same time?

- A. less than $\frac{1}{4}$ B. about $\frac{1}{2}$ C. about $\frac{3}{4}$ D. almost all

Question #4: Rating (1 2 3 4 5)

About how much of Earth's surface is covered by oceans?

- A. less than 20% B. about 50%
C. about 70% D. more than 90%

Question #5: Rating (1 2 3 4 5)

From Earth, we see the sun in the day sky and other stars in the night sky. Nighttime stars look like tiny points of light.

Which statement explains why nighttime stars appear so much smaller than the sun?

The stars are much smaller.

The sky is much darker at night.

The stars are much farther away.

The moon blocks out most starlight.

Question #6: Rating (1 2 3 4 5)

The weather forecast says a heavy snowstorm is coming later today.

Which weather observation is likely just before the snow?






- A. clear sky
- B. thick gray clouds
- C. small white clouds
- D. warm temperature

Question #7: Rating (1 2 3 4 5)

What causes day and night on Earth?

- A. tilting of Earth's axis
- B. rotation of Earth on its axis
- C. movement of Earth around the sun
- D. movement of the sun around Earth

A teacher gives students five rock samples to describe and sketch. The students record their observations in the table below. Use the information in the table to answer questions 8 – 10.

Sample	Observations	Sketch
Limestone	<ul style="list-style-type: none"> • tiny grains arranged in layers • feels gritty • reddish tan or gray • has a fish fossil 	
Conglomerate	<ul style="list-style-type: none"> • small rocks and pebbles of different colors stuck together • feels lumpy 	
Obsidian	<ul style="list-style-type: none"> • looks like black glass • cannot see parts of other things • feels smooth 	
Pumice	<ul style="list-style-type: none"> • light gray • has tiny holes like a sponge • very lightweight • feels very rough 	
Granite	<ul style="list-style-type: none"> • tiny specks that are black, white and gray • specks are about the same size • feels rough 	

Question #8: Rating (1 2 3 4 5)

Sedimentary rocks have visible layers of small pieces of other rocks. Based on the information in the rock sample table, which is a sedimentary rock?

- A. pumice B. granite C. obsidian D. limestone

Question #9: Rating (1 2 3 4 5)

How is conglomerate different from the other rock samples?

- A. It contains pebbles. B. It has visible layers.
C. It has a rough texture. D. It is all the same color.

Question #10: Rating (1 2 3 4 5)

Mount Rushmore is in South Dakota. This statue was carved more than 60 years ago. The faces on this granite statue are slowly wearing away.

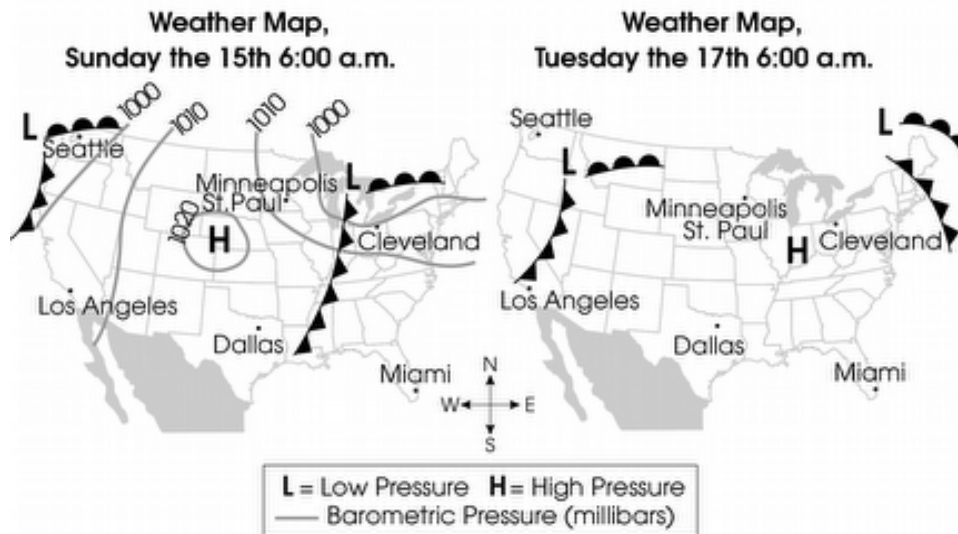
Which natural process causes this wearing away?

- A. earthquakes B. water from floods
C. blowing wind and rain D. lava flowing from volcanoes

Use the information in the table and maps below to answer questions 11 – 13.

The weather information shown below was reported on Sunday, the 15th of the month, and two days later on Tuesday, the 17th of the month. The table includes conditions for Sunday only, whereas the maps report early morning conditions for both Sunday and Tuesday.

Weather Conditions for Sunday at 6:00 a.m. and for Previous 24 Hours				
City	Previous 24 Hour Temperatures		Barometric Pressure 6:00 a.m. (millibars)	Relative Humidity 6:00 a.m. (percent)
	High (Fahrenheit)	Low (Fahrenheit)		
Cleveland	65	53	?	88
Dallas	75	50	1017	65
Los Angeles	68	50	1007	58
Miami	80	64	1016	60
Minneapolis	54	39	1007	65
Seattle	64	57	998	100



Question #11: Rating (1 2 3 4 5)

According to the weather map for Sunday, which is the approximate barometric pressure reading at Cleveland, Ohio, on Sunday at 6:00 a.m.?

- A. 990 millibars
- B. 995 millibars
- C. 1,000 millibars
- D. 1,010 millibars

Question #12: Rating (1 2 3 4 5)

Look at the weather map and the table for Sunday at 6:00 a.m. Fog was reported for one city on Sunday morning at 6:00 a.m. Which city was it?

- A. Dallas B. Miami C. Minneapolis D. Seattle

Question #13: Rating (1 2 3 4 5)

Using the two weather maps and the table of weather data above, predict the likelihood of precipitation and probable sky conditions (cloud cover) at Cleveland, Ohio, for Sunday and for the following Tuesday.

Give reasons for your predictions for each day.

Question #14: Rating (1 2 3 4 5)

Coal is usually found underground, compressed in a layer between other types of rock. Coal is produced by what rock-forming process?

- A. crystallization from melted rock
- B. formation of sediment from weathering
- C. deposition and burial of dead plant matter
- D. eruption of volcanic ash followed by settling in a layer

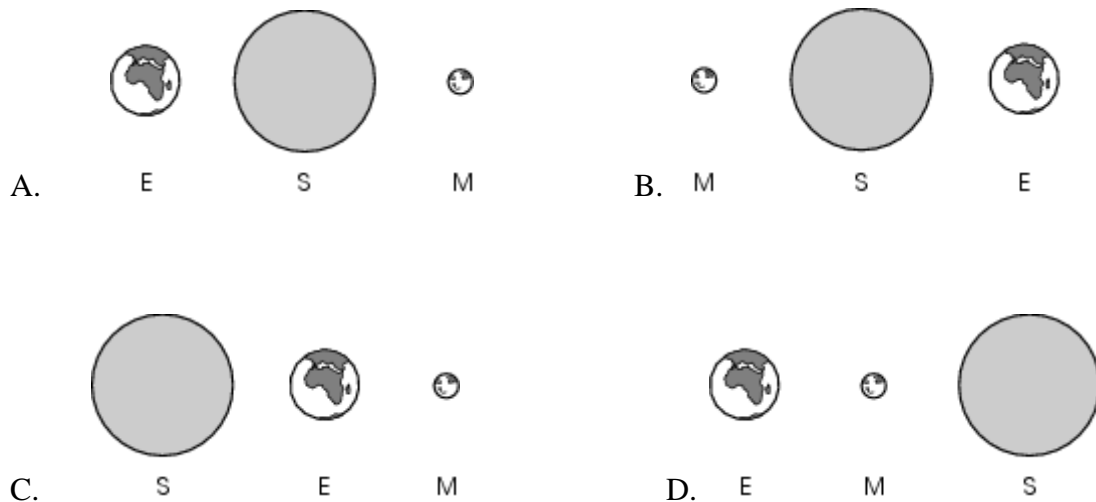
Question #15: Rating (1 2 3 4 5)

What is the major process of surface rock formation on volcanoes?

- A. Rock cools quickly from melted rock.
- B. Rock is recrystallized by extreme pressure.
- C. Rock solidifies slowly deep underground.
- D. Rock is formed from deposited sediment.

Question #16: Rating (1 2 3 4 5)

Which diagram shows the relative positions of Earth (E), the sun (S), and the moon (M) during a full moon? [**Note:** Diagrams are not drawn to scale.]

**Question #17: Rating (1 2 3 4 5)**

Why do satellites and spacecraft launched from Earth need to reach a specific speed to escape Earth's surface?

- A. to overcome Earth's gravitational force
- B. to protect equipment from radiation
- C. to break through the sound barrier
- D. to avoid Earth's magnetic field

Question #18: Rating (1 2 3 4 5)

This photograph of Galaxy M-82 was taken by students at the Kitt Peak observatory in Arizona.



Source: Adam Block; NOAO, AURA, NSF

Which type of equipment did the students use to collect the light to make this photograph?

- A. satellite B. binoculars C. microscope D. optical telescope

Question #19: Rating (1 2 3 4 5)

Which phenomena occur as a result of the gravitational attraction between the moon and Earth?

- A. Eclipses B. ocean tides
C. seasonal changes D. phases of the moon

Question #20: Rating (1 2 3 4 5)

What change would occur if Earth's rate of rotation significantly increased?

- A. The year would be shorter.
B. The year would be longer.
C. The day would be shorter.
D. The day would be longer.