

Running head: STUDENT VERBAL REPORTS AND EXPERT TEACHER EVALUATIONS

**An Experimental Test of Student Verbal Reports and Expert Teacher Evaluations as a Source of
Validity Evidence for Test Development**

Jacqueline P. Leighton

Colleen Heffernan

M. Kenneth Cor

Rebecca J. Gokiert

Ying Cui

Centre for Research in Applied Measurement and Evaluation (CRAME)

University of Alberta, Edmonton, CANADA

Acknowledgement: Preparation of this article was supported by a grant from the Social Sciences and Humanities Research Council of Canada (Grant No. 410-2003-0210). Grantees undertaking such projects are encouraged to express freely their professional judgment. This article, therefore, does not necessarily represent the positions or the policies of the Canadian government, and no official endorsement should be inferred. We are grateful to the Council of Ministers of Education, Canada for their help and responses to our questions about the SAIP Written Science Assessment. Moreover, we wish to thank the Edmonton Public School Board for making this research possible with their students and teachers.

Abstract

The *Standards for Educational and Psychological Testing* indicate that test instructions, and by extension item objectives, presented to examinees should be sufficiently clear and detailed to help ensure that they respond as developers intend them to respond (standard 3.20, AERA, APA, NCME, 1999). The present study investigates the use of verbal reports, one of many sources of evidence for validity arguments, as a way to inform the content clarity of 30 sample items from a large-scale science assessment. Student reports were used to edit items and create a *student-modified* test form. Evaluations were also solicited from expert teachers, which were used to edit the items and create an *expert-modified* test form. Both experimental forms, along with the original set of 30 items, were then randomly assigned to a sample of 264 examinees. Although hierarchical regression analyses indicated that examinee performance on the student-modified and expert-modified forms were similar in direction relative to performance on the original test items, item statistics revealed that student-modified test items were less discriminating than expert-modified test items. The implications of using verbal reports are discussed for informing test development.

An Experimental Test of Student Verbal Reports and Expert Teacher Evaluations as a Source of Validity Evidence for Test Development

Methods are constantly being sought to enhance test development and generate strong validity arguments. *Think-aloud protocol analysis* is one method that is increasingly considered as providing a source of evidence of the response processes measured by large-scale achievement test items. Think-aloud protocol analysis has origins in psychology (Ericsson & Simon, 1993) and involves interviewing participants individually as they respond to an item or task. During the interview, participants' verbalizations are captured using a video or audio recorder, and then transcribed verbatim to generate a *verbal report* for later analysis. Interviews often involve asking a participant to think aloud (a) *while* problem solving in order to gain insights of problem-solving processes or (b) *after* problem solving in order to gain insights of problem-solving beliefs about the nature of the task just solved (Chi, 1997; Willis, 2005). Verbal reports, as one of many sources of validity evidence, have been useful for researchers in their attempts to understand the response processes measured by traditional test item formats within large-scale assessments (e.g., Hamilton, Nussbaum, & Snow, 1997; Katz, Bennett, & Berger, 2000) and performance-based assessments (e.g., Ayala, Shavelson, Yin, & Schultz, 2002; Baxter & Glaser, 1998; Briggs, Alonzo, Schwab, & Wilson, 2006; see also the Standards for Educational and Psychological Testing, AERA, APA, NCME, 1999).

As a source of validity evidence, verbal reports can provide more information than just about the response processes examinees use to solve test items (Leighton, 2004; Mislevy, 2006). Verbal reports may also inform what Kane (2006a) refers to as the *content-related aspects* of a validity argument. In particular, verbal reports could be used to identify problems or ambiguities in the content clarity of items leading to the loss of comprehension for examinees. Item clarity is a necessary component of strong validity arguments since without it one could argue that inferences about what examinees know and can do are not premised on the most basic of assumptions—that the student understood what he or she was being asked to do on the test. Maximizing an examinee's comprehension of the content, including background material and instructions, pertaining to an item is basic and therefore often assumed when generating test-based inferences about examinee's knowledge and skills (Downing & Haladyna, 1997; Haladyna, 2004). In fact, in the Standards for Educational and Psychological Testing, standard 3.20 indicates that “the instructions

presented to test takers should contain sufficient detail so that test takers can respond to a task in the manner that the test developer intended” (p. 47). Item content clarity is of special interest given that examinees have been found to sometimes misinterpret the content of test items leading to the misalignment between intended targets of measurement and examinee response processes (Aikenhead, 1988; Aikenhead & Ryan, 1992; Ferrara, Duncan, Perie, Freed, McGivern, & Chilukuri, 2003; Ferrara et al., 2004; Norris, 1988). One reason why examinees may sometimes misinterpret test items is that items contain sources of ambiguity that test developers have missed in the refinement process.

Misalignment arising from item ambiguity poses a threat to validity arguments and test-based inferences (Cronbach & Meehl, 1955; Kane, 2006b; Messick, 1989) because an examinee’s test score may not be representative of the knowledge and skills learned. Although the practice of collecting verbal reports to inform item content clarity is not often undertaken, Haladyna (1997, p.233) states that “students can have brilliant insights, often pointing out flaws in items...[A]sking students for commentary promotes critical thinking about what is taught and what is tested. Students are the best sources for this analysis and criticism.” Using student reports to inform item content clarity along two dimensions—item comprehensibility and appearance—may be an important part of improving test development practices and securing evidence for validity arguments (Kane, 2006a, 2006b).

Conventional large-scale achievement test development practices do not often include the collection of verbal reports from examinees (Schmeiser & Welch, 2006). Instead, committees of content experts and technical specialists (e.g., psychometricians) decide on the design of the test, including philosophy, purpose and uses, intended examinee populations, administrative constraints, legal considerations, validity evidence, and test specifications or blueprint (Schmeiser & Welch, 2006). Test development (especially the design of multiple-choice items) is a labor-intensive and lengthy process, requiring item writers to undergo extensive training and to submit their items and stimuli for content, fairness, and measurement criteria approval. Although test items are field-tested with intended examinee populations, statistical indices (e.g., classical test theory or IRT item difficulty and discrimination) are typically the information gleaned from these field-tests (see Schmeiser & Welch, 2006). Student feedback in the form of verbal reports is not often considered or collected. While it seems prudent to withhold any recommendation to extend this already lengthy

development process by adding yet another source of data such as verbal reports, collecting verbal reports to gauge how well students understand items could bolster the development of strong validity arguments.

The objective of the present paper is to investigate whether student verbal reports, in comparison to expert judgments, are useful for informing item content clarity along two dimensions—item comprehensibility and appearance. The paper is divided into four sections: the first section provides a brief background into verbal reports and how different types of reports can be used to elucidate and inform distinct aspects of test item development for validity arguments; the second section describes the collection of (a) verbal reports from 54 students in response to 30 sample items from a large-scale achievement test in science, and (b) evaluations from two pre-service teachers with expertise in science assessment. This second section also describes how student verbal reports and expert evaluations were used to modify the 30 sample test items in order to create two experimental test forms to be tested with a new sample of examinees; the third section describes the results associated with administering the experimental forms to a new sample of 264 examinees, including examinee test performance and item statistics; the fourth section presents a discussion of the results, including the implications for using student verbal reports to inform test item development and validity arguments.

Section 1: Rationale for Using Verbal Reports for Test Item Development

Although verbal reports have been popularized as a method to identify examinees' response processes as they solve test items (e.g., Ferrara et al., 2003; Ferrara et al., 2004; Hamilton et al., 1997; Kane, 2006b; Katz et al., 2000; Norris, 1988; Snow & Lohman, 1989), this method could also be used to inform questions about item content clarity (Chi, 1997; Ericsson, 2006; Ericsson & Simon, 1993; Willis, 2005). In other words, verbal reports could provide evidence to satisfy different aspects of a validity argument—not just about examinees' response processes but also about item content clarity. Verbal reports are normally collected by conducting two types of interviews. The *concurrent interview* is used to gather evidence about the knowledge and skills a participant uses “in real time” as he or she is engaged in trying to find a solution to a task. The *retrospective interview* can be used to gather evidence about two different aspects of individual cognition—(a) domain-specific knowledge and skills and/or (b) metacognitive skills. When using the retrospective interview to gather evidence of metacognitive skills, the participant is asked to answer leading questions such as *why they chose a particular answer, how they think they solved the task, what they think*

about the task in terms of difficulty or ease, or why they decided to use a particular strategy to arrive at the answer (Chi, 1997; Taylor & Dionne, 2000). The information contained in participants' answers to these leading questions may be useful for understanding how they make sense of tasks and whether obstructions to item content clarity exist (Chi, 1997; Taylor & Dionne, 2000; Willis, 2005). This kind of evidence could be useful to test developers who are interested in gathering evidence, from the perspective of test-takers, about item clarity and in support of validity arguments. In the next section, we describe the methods used to collect evaluations from 54 grade 8 and grade 11 students, and two expert pre-service teachers in response to 30 sample items from a standardized large-scale science assessment. In this section we also describe the results of the student and teacher evaluations, and how they were used to modify 30 sample items in order to create two new experimental test forms.

Section 2: Method

Item Evaluation

Participants. Fifty-four moderate-ability students (14 girls and 16 boys in Grade 8; 16 girls and 8 boys in Grade 11) were selected from four science classes to provide item evaluations of 30 sample science items from the School Achievement Indicators Program (SAIP) 1999 Science Written Assessment (Council of Ministers of Education [CMEC], 2000). Moderate-ability students were recruited from two suburban junior-high schools and two suburban high schools willing to participate in the present study. Moderate-ability students were recruited because we expected these students to produce the most informative evaluations of potential problems with item clarity. We reasoned that high-ability students might not identify any items as ambiguous because they might know the subject material tested sufficiently well that even if some ambiguity existed, it might go unnoticed, and they would respond correctly. In contrast, low ability students might find many if not most items ambiguous because they would not know the material well. Their item evaluations might therefore be expected to reflect their lack of mastery and not a lack of item clarity. Hence, moderate-ability students were selected because they were expected to be able to identify problematic features in items without compensating for these features or confusing these features with lack of subject mastery. Participating junior high students had average science grades in the 60-80 percent range. Participating high school students had average science grades in the high 60-70 percent range. Two pre-service teachers with expertise in test development and classroom assessment were also involved in evaluating the 30 science test

items. The two teachers had specialization in science and had recently completed a test development/classroom-assessment course and had achieved top-standing in the course.

Materials: School Achievement Indicator Program (SAIP) Sample Science Test Items. The 30 sample items came from the 1999 SAIP Science Written Assessment, which is administered by the Council of Ministers of Education to students in both Grade 8 and Grade 11 (13- and 16-year-olds) every three to five years. This assessment is used to gauge students' comprehensive knowledge and problem solving in science. The SAIP Science Assessment includes multiple-choice (MC) and constructed-response (CR) test items (both dichotomously scored) representing three broad content domains: (a) knowledge and concepts of science (including knowledge in biology, chemistry, earth sciences, and physics), (b) nature of science, and (c) relationship of science to technology and societal issues. SAIP test items can also be categorized into three skill domains: conceptual (e.g., explain or define concepts, identify suitable examples of concepts), procedural (e.g., recognize when a procedure should be used, suggest procedures to solve particular problems), and use (e.g., formulate problems, apply a variety of strategies to solve problems, produce solutions, assess given solutions). It is critical to note that the items used in the present study were *sample* items not included in the actual administration of SAIP 1999 Science Assessment. These 30 items were not included in the actual administration because they represented additional and/or redundant items. Consequently, these 30 items were *expected to contain minor sources of ambiguity* given that they were not subjected to the rigorous item analysis that actual test items undergo for inclusion in the final SAIP Science Assessment. The 30 sample items consisted of 19 MC and 11 CR items.

Procedure. Each of the 54 students was interviewed individually in a quiet room for approximately 45 minutes using standard think-aloud instructions (Ericsson & Simon, 1993). During the interview, students were presented with approximately five items representing an *item set*; that is, the test items presented to students were contextually associated or linked to a narrative story and/or a specific background stimulus. Six different item sets were identified within the set of 30 sample items and students were randomly assigned to complete one of the six sets. A total of nine students completed each set (54 students divided by six sets = nine students per set). Although students were requested to "think aloud" concurrently and retrospectively, the retrospective portion of the interview was of primary interest. One retrospective question in particular was used to elicit student evaluations of item ambiguity: (a) *Imagine a student, like yourself, in your class. Do*

you think he or she might not understand this question? (b) How do you know this? Student responses during the interviews were audio-recorded and then transcribed verbatim by an independent, professional transcriber. Student-identified ambiguities of item clarity were then categorized according to Table 1 as either *structural/format* ambiguities (e.g., inconsistencies in item appearance that could cue test wiseness) or *contextual* ambiguities leading to a lack of item comprehensibility (Haladyna, 1997, 2004). More specifically, contextual ambiguities included deficiencies in the semantic framework of the item such as unnecessary or missing information in the target background material or in the item stem. Structural or format ambiguities were more mundane, pertaining largely to the appearance of the MC item stems and response options and the way in which information was organized in CR items.

Two graduate student assistants with formal knowledge of classroom assessment and item development worked together to classify the student evaluations of item ambiguity. Thirty-eight out of 54 students (21 grade 8 students and 17 grade 11 students) identified ambiguities in at least one item from the set of five they evaluated. Sixteen students did not identify ambiguities in any of the items. Tables 2 and 3 show a listing of the ambiguities identified by the 38 students in the MC and CR items, respectively. Of the identified ambiguities, 21.4% were vague and could not be interpreted as either contextual or structural (code of 00). Of the student-identified ambiguities in MC items that were clearly articulated, 67% were classified as contextual, with the remaining 33% classified as structural. One-hundred percent of student-identified CR item ambiguities were classified as contextual.¹

In addition to the student evaluations, two expert pre-service teachers used the categories in Table 1 to independently evaluate the 30 sample items. The teachers identified ambiguities in all 30 items. For expert-identified ambiguities, 51% of the MC item ambiguities were contextual. The remaining 49% of the MC item ambiguities were structural. For CR items, 84% were contextual and 16% were structural. The inter-rater agreement between the two expert teachers was .94 for item context and .95 for item structure.

Item Modification and Administration of Experimental Forms

Two graduate-student assistants made modifications to the 30 items based on the ambiguities identified by grade 8 and grade 11 students. Items were revised for clarity based strictly on ambiguities that would improve clarity without altering the item construct measured. For example, construct-relevant words or phrases were not modified. Only if a non-construct related word or phrase was identified as ambiguous,

was the word or phrase changed. If students identified a construct-relevant aspect of the item as difficult and ambiguous, this aspect of the item was kept intact so as to preserve the item construct. Twenty-six items were available for modification according to student-identified ambiguities, subject to the condition that the modifications not lead to a change in the item construct. Every item modification was cross-referenced to the student(s) that had identified the item ambiguity (see footnote 1). The 26 revised items along with the four items left unrevised were compiled into a new test booklet labeled *student-modified*.

The two graduate-student assistants also made modifications to the 30 items based on the ambiguities identified by the expert pre-service teachers. As with the student-based modifications, items were only modified according to expert-identified ambiguities if the alteration did not lead to changes in the item construct. Each revision was cross-referenced to the expert who identified the item ambiguity. Items modified according to expert-identified ambiguities were compiled into a new test booklet labeled *expert-modified*. The objective of developing these two new 30-item booklets (i.e., *student-modified* test form and *expert-modified* test form) was to administer them, along with the original, unmodified test items, to a new sample of grade 8 and grade 11 students.

Participants. A new sample of 156 grade 8 students (76 boys, 79 girls, 1 unidentified) and 108 grade 11 students (51 boys and 57 girls) from two middle-class suburban junior high and two middle-class suburban senior high schools were recruited and selected to complete one of three test forms: the original, unmodified form, student-modified form, or expert-modified form. The boys and girls in our sample had equivalent background mastery in science achievement as evidenced by the average of their last science report card mark, which was 76.6 (SD=12.26) and 77.4 (SD=12.46), respectively. The schools from which the students were sampled served a largely Caucasian population with the largest minority/ethnic group being South Asian. However, the majority of participating students in the study were Caucasian.

Procedure. Students were randomly assigned to complete one of three test forms: the original, unmodified form, the student-modified form, or the expert-modified form. Students were blind to the test form conditions. A similar front cover was used on all booklets, disguising the different conditions. Test administration conditions were standardized so that students completed test booklets independently at their desk during an interval of approximately 30 minutes. The 19 multiple-choice (MC) items and 11 constructed-response (CR) items were scored dichotomously using a scoring key obtained from the test developer (CMEC,

2000); a score of 0 was assigned to incorrect MC responses and a score of 1 was assigned to correct responses. CR answers had to be completely correct to obtain a score of 1; otherwise a score of 0 was assigned. The MC section of the form contained a total of 19 possible marks and the CR section of the form contained a total of 11 possible marks. Following the completion of test forms, a pre-service teacher with expertise in the science curriculum who was blind to the purpose of the study and the different test form conditions graded all 264 test forms.

Analysis

The first objective was to determine whether item modifications originating from expert evaluations and/or student verbal reports led to changes in the way examinees interpreted item objectives, and ultimately in the direction of their final scores. In particular, we expected modified items to be easier for examinees to answer (relative to the original, unmodified items) if construct-irrelevant cues in the original items had obscured the item objective. Likewise, we expected modified items to be more difficult for examinees to answer (relative to the original, unmodified items) if construct-irrelevant cues in the original items had been “giving away” the answer to the item. After computing descriptive statistics for each test form by item format, grade level, and gender, our first analysis involved hierarchical linear regression. Hierarchical linear regression analyses were conducted instead of factorial ANOVA because we wanted to determine the effect of test form on examinees’ MC scores and CR scores after controlling for known variables such as examinees’ last reported science mark, which was a continuous variable, grade level (grade 8=1, grade 11=2), and gender (male=1, female=2). After controlling for these variables in step 1 of the regression analysis, test form was entered into the regression analyses in step 2 as two variables using contrast coding; the first variable contrasted the original-unmodified form and the student-modified form (original-unrevised=1, student-modified=-1) and the second variable contrasted the student-modified form and the expert-modified form (student-modified=1, expert-modified=-1). We verified our results using ANOVA and report partial eta squares—an index of the effect sizes—for test form in the Results section.

The second analysis involved computing item statistics for the test items in both the expert-modified and student-modified test forms, and comparing these statistics with those of the original, unmodified test form. We computed item difficulty (p-values), item discrimination (point-biserial) values, and Cronbach’s alpha across the three test forms to help us interpret examinee performance differences found in the first

analysis, and determine the usefulness of the item modifications. In particular, item modifications leading to higher item discrimination values relative to the original, unmodified items were interpreted as being more useful than those revisions leading to comparable or lower item discrimination values relative to the original, unmodified items.

Section 3: Results

Examinee Performance

Shown in Tables 4 and 5 are examinee scores on the MC and CR sections of the SAIP assessment, respectively. A maximum score of 19 could be achieved on the MC section and a maximum score of 11 could be achieved on the CR section. The terms “higher” and “lower” will be used in this section to describe the relative magnitude of student scores in different test form conditions. It is important to note that the use of these terms (i.e., higher, lower) is not intended to imply any value judgment about the desirability of student performance. We recognize that whether a student scores well (higher, better) or not on a test does not necessarily indicate the adequacy of a test for measuring a construct. The adequacy of a test needs to be judged using criteria such as item statistics (discussed in the next section). However, we do think that efforts to improve test items should lead to changes in examinee test performance; otherwise there would be no purpose in refining items, either by removing construct-irrelevant sources of difficulty or easiness. As can be seen in Table 4, examinees in Grades 8 and 11 earned higher MC scores on the original, unmodified form relative to their scores on the student-modified and expert-modified forms. An inspection of Table 5 shows the opposite trend—examinees earned lower CR scores on the original, unmodified form relative to their performance on the student-modified and expert-modified forms. These descriptive results were subjected to further scrutiny by conducting hierarchical regression analyses.

Hierarchical regression results are reported separately for MC and CR scores in Tables 6 and 7, respectively. Shown in Table 6 is the analysis of MC scores. The first step of the analysis indicated that expected variables such as grade level and last report-card mark in science significantly predicted MC scores. In other words, grade 11 students earned higher MC scores relative to grade 8 students across all three test forms ($\beta=0.218^{**}$) and those students with higher report-card marks in science also earned higher MC scores than those with lower report-card marks in science ($\beta=.434^{**}$). Gender had no association to MC scores.² At the second step of the analysis, there were significant differences in examinee performance by test form. In

particular, those examinees assigned to complete the original, unmodified test form earned higher MC scores than those examinees assigned to complete the student-modified form ($\beta=0.271^{**}$). In turn, examinees assigned to complete the student-modified test form earned higher MC scores than those assigned to complete the expert-modified form ($\beta=0.277^{**}$). The upshot of these results is that even once grade level and previous science mark were held constant, examinees assigned to the original, unmodified test condition scored approximately 1 unit higher (weighted $\bar{x}=12.45$ out of 19 or 66%) on the MC section of the test than those students assigned to the student-modified condition (weighted $\bar{x}=11.59$ out of 19 or 61%). In addition, examinees assigned to the student-modified condition scored approximately 1 unit higher on the MC section of the test than those students assigned to the expert-modified condition (weighted $\bar{x}=10.62$ out of 19 or 56%). A partial $\eta^2=0.094$ was calculated for the effect of test form instead of a classic η^2 because we wanted to evaluate this effect only against the remaining variance unaccounted for by other controlled variables in the study (see Cohen, 1973, p. 110). A value of 0.094 is considered to be of moderate magnitude in the behavioral and social sciences (Cohen, 1988).

Shown in Table 7 is the analysis of CR scores. Similar to the MC analyses, grade level and last report-card mark in science predicted CR scores (gender predicted CR scores at step 1 of the analysis, but this variable was no longer significant at step 2). There were also significant differences in examinee performance on CR items by test form. Examinees assigned to complete the expert-modified test form earned higher CR scores relative to examinees assigned to complete the student-modified form ($\beta=-0.176^{**}$). Examinees earned the lowest CR scores on the original, unmodified items ($\beta=-0.285^{**}$). The net significance of these results is that even once grade level and previous science mark are held constant, examinees assigned to the original, unmodified test condition scored close to a unit lower (weighted $\bar{x}=5.72$ out of 11 or 52%) than those students assigned to the student-modified condition (weighted $\bar{x}=6.67$ out of 11 or 61%), and examinees assigned to the student-modified condition scored lower again relative to those students assigned to the expert-modified condition (6.84 out of 11 or 62%). A check of these results using ANOVA revealed a partial η^2 of 0.088 for the effect of test form.

Item Performance

From the examinee performance described in the previous section, it would appear as if the student-based modifications were similar in nature to expert-based modifications. Examinees scored lower on both the expert-modified and student-modified MC items compared to the original, unmodified MC items. Likewise, examinees scored higher on both expert-modified and student-modified CR items than on the original, unmodified items. A classical test theory item analyses was conducted to check on any differences found among test forms collapsed across grade levels. We collapsed across grade levels for two reasons: first, grade level was not shown to interact with test form in examinee test performance. Second, item statistics for SAIP are routinely computed across grade level.

Reliabilities were calculated for the three test forms. Cronbach's alpha was 0.698, 0.626, and 0.582 for the original-unmodified, student-modified, and expert-modified test forms, respectively. Interestingly, reliabilities decreased for the experimental test forms even though they were modified precisely to improve test items. We also computed item p-values across the three test forms. These values are shown in Table 8. For MC items, the original, unmodified form had the highest average p-value (0.654), followed by the student-modified form (0.610), and then the expert-modified form (0.560). For CR items, the original, unmodified form had the lowest average p-value (0.518), followed by the student-modified form (0.607), and then the expert-modified form (0.622). A repeated-measures analysis revealed there were no differences in item difficulty across the three test forms; however, there was a significant interaction between test form and item format, $F(2, 56)=5.35, p<.01$, partial $\eta^2=0.16$.³ Expert- and student-modified MC items were generally more difficult for students than the original, unmodified MC items, whereas expert- and student-modified CR items were generally easier for students than the original, unmodified CR items.

We computed item point-biserial values across the three test forms. These values are shown in Table 9. For MC items, the expert-modified form had the highest average point-biserial value (0.270), followed by the original, unmodified form (0.227), and then the student-modified form (0.188). For CR items, the expert-modified form had the highest average point-biserial value (0.267), followed by the original, unmodified form (0.233), and then the student-modified form (0.188). A repeated-measures analysis revealed a statistically significant difference in item discrimination, $F(2,56)=3.584, p<.05$, partial $\eta^2=0.11$ (see footnote 3) across the three test forms; however, the largest difference was between the expert-modified and the student-

modified form. There was no significant interaction effect between test form and item format for point-biserial values.

Analysis of changes for MC. Although documenting the line-by-line modifications made to all items based on expert and student comments would exceed the limits of this section, this information is available by contacting the first author. In this section, we illustrate examples of student-based and expert-based modifications made to science items, which led to increases in point-biserial values relative to the original, unmodified items (see footnote 1). We focus on increases in point-biserial values, instead of item difficulty values, as a source of evidence for item improvements because item difficulty values can increase for construct-irrelevant reasons. In contrast, increases in item discrimination are usually desirable for improvements in item quality. Inspection of Table 9 reveals that expert-based modifications led to improvements in point-biserial values for 11 out of 19 MC items (items 4, 9, 10, 11, 13, 17, 20, 22, 23, 29, 30). Of these 11 items with improved point-biserial values, 9 items became more difficult (lower p-values) with the exception of items 22 and 23. Student-based modifications led to improvements in point-biserial values for 6 out of 19 items (items 11, 13, 17, 23, 27, and 30). However, the increase in point-biserial values for the student-modified items was usually of a smaller magnitude in comparison to the expert-modified items. Of the 6 student-modified items with improved point-biserial values, 3 of them became more difficult relative to the original, unmodified items (items 13, 17, and 30), and 3 became less difficult in relation to the original, unmodified items (items 11, 23, and 27). (Although item 10 shows an improvement in point-biserial value, this is due to chance since student comments for this item were unable to be interpreted, as shown in Table 2, and therefore did not lead to modification of the item.)

There were 5 common items (items 11, 13, 17, 23, and 30) for which both student- and expert-based modifications led to improvements in point-biserial values relative to the original item values. Three of these modified items led to increases in item difficulty (items 13, 17, and 30) with the exception of item 23, which became easier on both the student-modified and expert-modified form, and item 11 which became easier on the student-modified form but not the expert-modified form. What kind of comments made by experts and students might have led to improvements in these point-biserial values? An illustrative example is item 17. The original item 17 had a difficulty level of 0.789 and a point-biserial of 0.234. Experts indicated that this item had:

- alternatives that were not consistent in content (some were general and others too specific),
- alternatives were written so there could be more than one correct answer,
- distracters that were not all plausible,
- a stem that contained unnecessary information, and
- background material that was irrelevant to answering the item (item could be answered without background material).

Likewise, students who reported that this item could be misunderstood by a classmate said that this item had:

- responses that were not homogeneous in content
- words of low frequency that are unclear and not related to the item construct
- words of low frequency that are scientific and related to the item construct (e.g., word “species”)
- a stem that lacked necessary information for an informed student to answer the item correctly.

A comparison of these comments suggests that experts were generally more precise than students in pointing out the parts of the item that were problematic by zeroing in on features such as indicating that some of the distracters were implausible. Furthermore, as shown in students’ evaluation of item 17, in some cases students identified words as problematic but these evaluations did not lead to changes in the item because the scientific words were associated with the item construct. However, both experts and students were alike in identifying when the item stem lacked the adequate information for answering the question. When expert-based and student-based modifications were made to item 17, point-biserial values rose to 0.591 and 0.257, respectively (original was 0.234). Furthermore, item p-values decreased to 0.412 and 0.449, respectively (original was 0.789). Thus, item 17 became more difficult after making the modifications but its discrimination improved.

Analysis of Changes for CR. Unlike the modifications to MC items, which generally led to increases in item difficulty, modifications to CR items typically lowered difficulty in relation to the original, unmodified items. An analysis of Table 9 reveals that expert-based modifications led to improvements in point-biserial values for 7 out of 11 CR items (items 2, 6, 7, 16, 19, 21, and 24). In all but 3 cases (items 2, 21, and 24), the modified CR items became easier (i.e., led to higher p-values). Student-based modifications led to

improvements in point-biserial values for 4 out of 11 CR items (items 15, 19, 21, and 24). Of these 4 items, 3 became easier (higher p-values) and 1 became negligibly more difficult for students to answer (item 24). There were only 3 common items for which both expert- and student-modifications led to improvements in point-biserial values (items 19, 21, and 24). Of these 3 items, only two had comparable item difficulties in terms of becoming easier or more difficult relative to the original items (items 19 and 24).

What kind of comments made by experts and students lead to improvements in point-biserial values for CR items? An example is item 19. The original item had a difficulty level of 0.622 and point-biserial value of 0.185. Experts commented that this item was problematic because *the background material was unnecessary to answer the question*. Students indicated that this item was confusing because *the stem lacked necessary information for a knowledgeable student to answer the item correctly* (e.g., units of measurement were not consistent in question and conversion of units was part of the learning outcomes). A comparison of these comments suggests that experts were concerned about the quality of information contained in the background story and its possible irrelevance to the item of interest whereas students were concerned about consistency in specific contextual features related to producing the correct response. When expert-based and student-based modifications were made to item 19 based on their comments, point-biserial values rose to 0.223 and 0.220, respectively (original was 0.185). Furthermore, item p-values increased to 0.741 and 0.764, respectively (original was 0.622). Thus, item 19 became easier to solve and more discriminating after modifying it.

Section 4: Summary and Discussion

The objective of the this study was to investigate whether student verbal reports, in comparison to expert teacher judgments, are useful for informing item content clarity and, by extension, adding useful evidence to the generation of validity arguments. To address this objective, 54 grade 8 and 11 students of moderate ability were recruited to engage in think-aloud interviews. During the retrospective portion of the interview, students were asked whether any of the items they solved might be misunderstood by a classmate and for what reasons. Two pre-service teachers with expertise in science assessment also evaluated the 30 sample items using a list of well-known assessment design principles (see Table 1). Following the student and teacher evaluations, the 30 test items were modified according to student-identified ambiguities and expert-identified ambiguities to create two new test forms—a student-modified test form and an expert-modified

test form. These two test forms, along with the original, unmodified 30 test items, were then administered to a new sample of Grade 8 and 11 students to determine how different sources of item revisions compared in terms of influencing examinee and item performance. The results of the hierarchical regression analyses revealed that the student-modified and expert-modified test forms improved performance among examinees for CR items but not for MC items in relation to the original, unmodified test form. One perspective on this result is that modified MC items became more difficult relative to the original, unmodified MC items because this item type was more amenable to structural modifications (see Table 1 and section Item Evaluation), which had the potential to alleviate test-wiseness and thus make the items more challenging. In contrast, the modified CR items became less difficult relative to the original, unmodified CR items because the modifications of this item type largely pertained to infusing more contextual information in the item stems.

Moreover, the results of the item analyses revealed that while expert- and student-modified items had comparable item-p values (see Table 8), the expert-modified items more often led to improvement in item discrimination and had, on average, higher item discrimination values than the student-modified items. However, it is also constructive to consider that the reliability of the expert-modified test form (0.582) decreased relative to the original, unmodified form (0.698) and student-modified test form (0.626). Although an explanation of this decrease is speculative at this stage, it has been suggested by Moss (1994) that improvements in the measures of constructs could have an adverse effect on reliability, in particular internal consistency estimates or Cronbach's alpha (see Mislevy, 2004). This adverse effect could occur when multiple constructs underlie test items and a formula that is premised on the inter-relatedness of a set of items is used to calculate reliability (see Hattie, 1985; Schmitt, 1996). Notwithstanding this outcome in reliability for the expert-modified test form, increases in item discrimination values are generally desirable for test development and, consequently, for validity arguments. Thus, the results of this study suggest that expert pre-service teachers provided more useful information than students for item development purposes.

At face value this conclusion suggests that student verbal reports may not be as informative for test development and validity arguments as has been heretofore assumed. However, there are at least three issues to consider when interpreting this conclusion. First, the 54 students who provided item evaluations were of moderate-ability and may not have been sufficiently proficient at recognizing or articulating specific sources of item ambiguity. In the absence of specific comments about the source of item ambiguity, items

could not be modified. Consider that some segments of student verbal reports were coded as *not interpretable*, meaning that although students identified the item as potentially confusing they did not identify a specific source for the ambiguity; hence, no modifications could be made. The segments that led to a classification of *not interpretable* included student comments such as “answers are a bit confusing” or “this is a confusing kind of statement” or “question is poorly written and could be read incorrectly” without any specific information to guide focused item modification. There is little doubt that these comments are potentially quite valuable but in these cases students were unable to articulate precisely what part of the answers, stems, or statements were causing problems—even after students were probed with the question “*How do you know this?*” (see section Item Evaluation). To avoid this situation, we could have interviewed students of higher average ability to provide items evaluations. However, high-ability students can introduce different types of challenges to verbal report studies because they are highly proficient within the domain being tested. One challenge with involving high-ability students in think-aloud studies is their ability to respond to moderately-difficult tasks quickly using recall (Leighton, 2004). Many of these students have automated considerable knowledge and skills, and could be expected to compensate for potential item ambiguities by filling in gaps with appropriate information that other, less sophisticated students may not be able to do. Nevertheless, one line of research we have initiated is to investigate the verbal reports of higher-achieving students to determine whether these students are better able to recognize sources of item ambiguity and other problematic features (Gokiert & Leighton, 2008).

Assuming that students of moderate-ability were an appropriate group to interview for item evaluations, it is possible that when they did provide specific item evaluations, these evaluations were not interpreted or translated effectively by the graduate research assistants. While this is possible with any interpretative exercise, it is unlikely in this case. If the translation process had any bias built into it, it would have been for the graduate student assistants to infuse more meaning into the student evaluations rather than less. Given the graduate assistants’ knowledge of item development, the bias would have been to make the student evaluations *more* meaningful rather than less meaningful, thus leading to more revisions of problematic features in the student-modified test form; a finding that was not observed as evidenced by the item discrimination values. The student evaluations generally lead to less discriminating items even compared to the original, unmodified items. However, potential bias should be viewed as a note of caution for

using student verbal reports in test development and related validity arguments. While student verbal reports need to be interpreted by knowledgeable individuals who know how to revise identified ambiguities, it is critical to not imbue what students say with more meaning than is originally intended.

The third issue to consider is that the students who provided verbal reports were at a practical disadvantage relative to our expert pre-service teachers. The students we interviewed provided item evaluations in the absence of the list shown in Table 1. This list outlines major categories of item features to consider when conducting an item evaluation. The expert pre-service teachers in our study were familiar with these categories and worked directly from them. Students were not provided with the list when we asked them if there was anything in the item that could be potentially misunderstood by a classmate. We did not provide them with the list because we wanted to determine if they could identify potential sources of ambiguity spontaneously without being guided by a list. Students who are selected to provide verbal reports are normally interviewed without scaffolds or supporting materials to influence their verbalizations. It is rare during think aloud interviews to have students work from a list or rubric in responding to a task because the interest is usually to measure the cognitive processes students have developed based on their learning histories rather than any process cued or learned on-the-spot during the interview. However, this may not be a concern when seeking student item evaluations because in this case we are not interested in measuring the underlying cognitive processes students have developed for solving items as much as we are interested in measuring something more basic—their basic comprehension of what the item is requesting. If we can facilitate student evaluations by providing materials with which they can engage more fully in the evaluation, this might bring us closer to our objective of understanding how clear test items are to test-takers. Some support for this conjecture can be found in studies of text comprehension, where 8- and 10-year-old who are children instructed on how to detect inconsistencies in text are better able to detect inconsistencies relative to a control group (Markman & Gorin, 1981; see also Rubman & Waters, 2000).

Thus, if verbal reports are collected from students to inform item development and validity arguments, it is important to recognize that spontaneous verbalizations from students may not be sufficiently specific to inform item modifications as well as those obtained from expert teachers. One line of future research would be to investigate whether students can be trained to use age-appropriate rubrics to evaluate test items and whether doing so leads to similar evaluations as those provided by experts. Alternatively, in

the absence of using a rubric, students could be asked directly to specify the parts of the item that should be changed (Gokiert & Leighton, 2008). However, in the absence of a rubric, the risk would always be that student knowledge was being underestimated due to the need to reconstruct potential problems instead of being able to recognize the problems on a list. As Markman and Gorin (1981, p. 325) state “when children are given examples of what types of problems to look for, they are capable of adjusting their standard of evaluation.” Age-appropriate rubrics could be created so that younger students could recognize and classify potentially problematic item features in familiar language. Students would need to be trained to use the rubric to evaluate the items. However, having done so, test developers would have better information of examinees’ perspectives on the items they were solving, which would strengthen validity arguments based in part on student interview data.

Footnotes

¹The full listing of evaluations made by experts and student verbal reports for all items are not reproduced here because of space limitations. However, they are available upon request from the first author.

²Although gender was found to be nonsignificant in the first step of the analysis, it was retained in the second step of the hierarchical regression analysis so as to determine its effect among the presence of other predictors (see Field, 2005).

³The results are the same even once item 8 is removed which appears to be an outlier in the data. This item yielded very different p-values within the student- and expert-modified test conditions. The point-biserial for this item was improved in the student-modified condition but not in the expert-modified condition. An analysis of modifications requested by experts for this item did not reveal any idiosyncracies that could shed light on the low discrimination index.

References

- Aikenhead, G.S. (1988). An analysis of four ways of assessing student beliefs about STS topics. *Journal of Research in Science Teaching*, 25, 607-629.
- Aikenhead, G.S., & Ryan, A.G. (1992). The development of a new instrument: "Views on science-technology-society" (VOSTS). *Science Education*, 76, 477-491.
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Author. Washington, DC.
- Ayala, C.C., Shavelson, R.J., Yin, Y., & Schultz, S.E. (2002). Reasoning dimensions underlying science achievement: The case of performance assessment. *Educational Assessment*, 8, 101-121.
- Baxter, G. P. & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice*, 17, 37-45.
- Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment*, 11(1), 33-63.
- Chi, M.T.H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences*, 6, 271-315.
- Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor and ANOVA designs. *Educational and Psychological Measurement*, 33, 107-112.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Council Ministers of Education, Canada (2000). *Public report on science assessment: SAIP School Achievement Indicators Program 1999*. Retrieved August 12, 2002, from <http://www.cmec.ca/saip/science2/science2.en.stm>.
- Cronbach, L.J., & Meehl, P. E. (1955). *Psychological Bulletin*, 52, 281-302.
- Downing, S.M., & Haladyna, T.M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education*, 10, 61-82.
- Ericsson, K.A. (2006). Protocol analysis and expert thought: concurrent verbalizations of thinking during experts' performance on representative tasks (pp. 223-241). In K.A. Ericsson, N. Charness, P.J. Feltovich,

- R.R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance*. Cambridge, UK: Cambridge University Press.
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol Analysis*. Cambridge, MA: The MIT Press.
- Ferrara, S., Duncan, T.G., Freed, R., Velez-Paschke, A., McGivern, J., Mushlin, S., Mattessich, A., Rogers, A., & Westphalen, K. (April, 2004). *Examining test score validity by examining item construct validity*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.
- Ferrara, S., Duncan, T., Perie, M., Freed, R., McGivern, J., & Chilukuri, R. (April, 2003). *Item construct validity: Early results from a study of the relationship between intended and actual cognitive demands in a middle school science assessment*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Field, A. (2005). *Discovering statistics using SPSS (2nd ed.)*. Sage.
- Gokiert, R.J., & Leighton, J.P. (2008, April). *Large-scale science assessment: Three forms of construct validity evidence*. Paper to be presented at the annual meeting of the National Council on Measurement in Education (NCME), New York City, New York.
- Gronlund, N.E. (2003). *Assessment of student achievement (7th ed.)*. Boston, MA: Allyn and Bacon.
- Haladyna, T.M. (1997). *Writing test items to evaluate higher-order thinking*. Boston: Allyn and Bacon.
- Haladyna, T.M. (2004). *Developing and validating multiple-choice test items (3rd ed.)*. Mahwah, NJ: Erlbaum.
- Hamilton, L.S., Nussbaum, E. M., & Snow, R. E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education, 10*, 181-200.
- Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139-164.
- Kane, M.T. (2006a). Content-related validity evidence. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development* (pp. 131-154). Mahwah, NJ: Erlbaum.
- Kane, M. T. (2006b). Validation. In R. L. Brennan (Ed.), *Educational measurement (4th ed., pp. 17-64)*. Westport, CT: National Council on Measurement in Education and American Council on Education.

- Katz, I. R., Bennett, E., & Berger, A. E. (2000). Effects of response format on difficulty of SAT-Mathematics items: It's not the strategy. *Journal of Educational Measurement, 37*, 39-57.
- Leighton, J. P. (2004). Avoiding Misconceptions, Misuse, and Missed Opportunities: The Collection of Verbal Reports in Educational Achievement Testing. *Educational Measurement: Issues and Practice, Winter*, 1-10.
- Markman, E.M., & Gorin, L. (1981). Children's ability to adjust their standards for evaluating comprehension. *Journal of Educational Psychology, 73*, 320-325.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 1-103). New York: American Council on Education/Macmillan.
- Mislevy, R. J. (2004). Can there be reliability without "reliability "? *Journal of Educational and Behavioral Statistics, 29*, 241-244.
- Mislevy, R.J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257-305). Westport, CT: National Council on Measurement in Education and American Council on Education.
- Moss, P.A. (1994). Can there be validity without reliability? *Educational Researcher, 23*, 5-12.
- Norris, S.P. (1988). Controlling for background beliefs when developing multiple-choice critical thinking tests. *Educational Measurement: Issues and Practice, 7*, 5-11.
- Rubman, C.N. & Waters, H.S. (2000). A, B seeing: The role of constructive processes in children's comprehension monitoring. *Journal of Educational Psychology, 92*, 503-514.
- Schmeiser, C.B. & Welch, C.J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307-353). Westport, CT: National Council on Measurement in Education and American Council on Education.
- Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*, 350-353.
- Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263-331). New York: American Council on Education, Macmillan.

Taylor, K. L., & Dionne, J-P. (2000). Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology, 92*, 413-425.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.

Table 1
Codes and Associated Content Categories for Evaluating Science Test Items

STRUCTURE	
CODE	
00	Indefinite
Test-wiseness	
01	Similar wording exists in the stem and at least one of the responses
02	Absolutes or specific determiners are used in the item (e.g., always, never, none, only)
03	There are grammatical clues (e.g., a instead of an) in the stem and/or responses
04	The length and/or detail of the correct answer is significantly greater than the distracters
05	The correct answer is stated in 'textbook' or stereotyped language that enables the uninformed student to select it
Formatting	
06	Item type is not the best method to assess the outcome
07	The standard multiple choice layout is not used (e.g., incomplete statement or question followed by responses)
08	There is repetitive wording in the responses that should be in the stem
09	Key words (e.g., best, main, negatives) are not emphasized by bolding, CAPS or by underlining
10	Units of measurement are not included in the item
Responses	
11	There is not one clearly correct answer in multiple choice responses
12	The distracters are not plausible and attractive to the uninformed (possible testwiseness)
13	The responses are not homogeneous (possible testwiseness)
14	The responses are not presented in a logical sequence e.g., alpha / numeric
Ambiguous Words	
14a	Low-frequency, unclear, and NOT associated with measured construct
14b	Low-frequency word, scientific, and related to measured construct
CONTEXT	
CODES	
Item Context	
15	The item contains visuals that are unnecessary or unclear
16	The item tests multiple concepts, skills or problems
17	The stem lacks necessary information that is required for an informed student to answer the item correctly
18	The stem includes nonfunctional/irrelevant information that may prevent the informed student from answering the item correctly
19	The item makes assumptions about prior knowledge or uses language that introduces bias toward a specific group
Background Material (Story) for Item Completion	
20	Background material is unnecessary to correctly answer the question
21	Background material is lengthy and unreadable
22	Background material is irrelevant to the learner outcome

Note. Contents developed from Gronlund (2003). Used to categorize student-identified ambiguities and used by expert teachers to evaluate multiple-choice and constructed-response test items in SAIP Science Assessment

Table 2

Student-Identified Ambiguities By Multiple Choice Item

ITEM	IDENTIFIED AMBIGUITIES
MC1:	18-The stem includes nonfunctional/irrelevant information that may prevent the informed student from answering the item correctly 20-Background material is unnecessary to correctly answer the question
MC3:	00- Not interpretable (2) 14b-Low-frequency word, scientific, and related to measured construct 18-The stem includes nonfunctional/irrelevant information that may prevent the informed student from answering the item correctly
MC4:	00- Not interpretable 17-The stem lacks necessary information that is required for an informed student to answer the item correctly
MC5:	14a-Low-frequency, unclear, and NOT associated with measured construct 17- The stem lacks necessary information that is required for an informed student to answer the item correctly (4) 20-Background material is unnecessary to correctly answer the question
MC8:	00-Not interpretable 01-Similar wording exists in the stem and at least one of the responses 11- There is not one clearly correct answer in multiple choice responses 18-The stem includes nonfunctional/irrelevant information that may prevent the informed student from answering the item correctly (2)
MC9:	11-There is not one clearly correct answer in multiple choice responses (2) 12- The distracters are not plausible and attractive to the uninformed 17- The stem lacks necessary information that is required for an informed student to answer the item correctly (2)
MC10:	00- Not interpretable (2)
MC11:	00- Not interpretable 18- The stem includes nonfunctional/irrelevant information that may prevent the informed student from answering the item correctly (3)
MC13:	00- Not interpretable 12- The distracters are not plausible and attractive to the uninformed 18- The stem includes nonfunctional/irrelevant information that may prevent the informed student from answering the item correctly
MC17:	13- The responses are not homogeneous 14a- Low-frequency, unclear, and NOT associated with measured construct 14b- Low-frequency word, scientific, and related to measured construct 17- The stem lacks necessary information that is required for an informed student to answer the item correctly
MC18:	07- The standard multiple choice layout is not used (e.g., incomplete statement or question followed by responses) 19- The item makes assumptions about prior knowledge or uses language that introduces bias toward a specific group
MC20:	20- Background material is unnecessary to correctly answer the question (2) 21- Background material is lengthy and unreadable (2)
MC22:	00- Not interpretable 19- The item makes assumptions about prior knowledge or uses language that introduces bias toward a specific group (2)
MC23:	14b- Low-frequency word, scientific, and related to measured construct 18- The stem includes nonfunctional/irrelevant information that may prevent the informed

	student from answering the item correctly
MC25:	13- The responses are not homogeneous 18- The stem includes nonfunctional/irrelevant information that may prevent the informed student from answering the item correctly
MC27:	11- There is not one clearly correct answer in multiple choice responses
MC28:	12- The distracters are not plausible and attractive to the uninformed 18- The stem includes nonfunctional/irrelevant information that may prevent the informed student from answering the item correctly (3)
MC29:	00- Not interpretable 11- There is not one clearly correct answer in multiple choice responses
MC30:	11- There is not one clearly correct answer in multiple choice responses 14b- Low-frequency word, scientific, and related to measured construct 15- The item contains visuals that are unnecessary or unclear

Note. Number in parentheses indicates the frequency this comment was made by different students

Table 3

Student-Identified Ambiguities By Constructed Response Item

ITEM	IDENTIFIED AMBIGUITIES
CR2:	00- Not interpretable 14a- Low-frequency, unclear, and NOT associated with measured construct 15- The item contains visuals that are unnecessary or unclear
CR6:	17-The stem lacks necessary information that is required for an informed student to answer the item correctly
CR7:	18- The stem includes nonfunctional/irrelevant information that may prevent the informed student from answering the item correctly 19- The item makes assumptions about prior knowledge or uses language that introduces bias toward a specific group
CR12:	00- Not interpretable 14a-Low-frequency, unclear, and NOT associated with measured construct (2) 18-The stem includes nonfunctional/irrelevant information that may prevent the informed student from answering the item correctly
CR14:	00-Not interpretable (4) 14a- Low-frequency, unclear, and NOT associated with measured construct (2) 17- The stem lacks necessary information that is required for an informed student to answer the item correctly 18- The stem includes nonfunctional/irrelevant information that may prevent the informed student from answering the item correctly (2)
CR15:	14a- Low-frequency, unclear, and NOT associated with measured construct 14b- Low-frequency word, scientific, and related to measured construct
CR16:	14b- Low-frequency word, scientific, and related to measured construct
CR19:	18- The stem includes nonfunctional/irrelevant information that may prevent the informed student from answering the item correctly
CR21:	00- Not interpretable (3) 17- The stem lacks necessary information that is required for an informed student to answer the item correctly 19- The item makes assumptions about prior knowledge or uses language that introduces bias toward a specific group (2)
CR24:	14a- Low-frequency, unclear, and NOT associated with measured construct 19- The item makes assumptions about prior knowledge or uses language that introduces bias toward a specific group (2)
CR26:	00- Not interpretable 18- The stem includes nonfunctional/irrelevant information that may prevent the informed student from answering the item correctly

Note. Number in parentheses indicates the frequency this comment was made by different students

Table 4

Mean Scores and Standard Deviations for Examinees' Performance on Multiple-Choice Items by Gender, Grade and Test Form

		Multiple Choice			
	Gender (<i>n</i>)	Original	Student-Modified	Expert-Modified	Total
Grade 8	Male (76)	11.26 (2.786)	10.71 (2.901)	9.54 (2.353)	10.45 (2.749)
	Female (79)	12.29 (2.652)	11.48 (2.421)	10.57 (2.233)	11.57 (2.550)
	Total (155)	11.93 (2.719)	11.02 (2.719)	10.02 (2.332)	11.02 (2.700)
Grade 11	Male (51)	13.88 (2.781)	12.35 (2.090)	11.47 (2.787)	12.57 (2.715)
	Female (57)	12.58 (3.339)	12.45 (2.704)	11.50 (2.728)	12.19 (2.924)
	Total (108)	13.19 (3.115)	12.41 (2.409)	11.49 (2.716)	12.37 (2.820)

Note. Maximum score on MC section is 19; Standard deviations are in parenthesis.

Table 5

Mean Scores and Standard Deviations for Examinees' Performance on Constructed-Response Items by Gender, Grade and Test Form

		Constructed Response			
	Gender (<i>n</i>)	Original	Student-Modified	Expert-Modified	Total
Grade 8	Male (76)	5.47 (1.172)	6.23 (1.606)	6.23 (1.608)	6.04 (1.527)
	Female (79)	4.77 (1.308)	5.81 (1.965)	6.35 (1.613)	5.51 (1.716)
	Total (155)	5.02 (1.296)	6.06 (1.754)	6.29 (1.594)	5.77 (1.643)
Grade 11	Male (51)	6.88 (2.176)	7.53 (1.419)	7.65 (1.618)	7.35 (1.764)
	Female (57)	6.58 (2.168)	7.55 (1.468)	7.61 (1.461)	7.25 (1.766)
	Total (108)	6.72 (2.146)	7.54 (1.426)	7.63 (1.516)	7.30 (1.758)

Note. Maximum score on CR section is 11; Standard deviations are in parenthesis.

Table 6

Summary of Hierarchical Regression Analysis for Variables Predicting Multiple-Choice Scores in Science Assessment (N=262)

Variable	B	SE B	β
Step 1			
Grade level	1.247	0.308	.218**
Science mark	0.099	0.012	.434**
Gender	0.426	0.304	.076
Step 2			
Grade level	1.269	0.294	.221**
Science mark	0.101	0.012	.443**
Gender	0.286	0.292	.051
Original vs. student	0.926	0.205	.271**
Student vs. expert	0.964	0.207	.277**

Note. $R^2=.251$ for Step 1; $R^2=.325$ for Step 2; $\Delta R^2=0.074$; * $p<.05$, ** $p<.01$

Table 7

Summary of Hierarchical Regression Analysis for Variables Predicting Constructed Response Scores in Science Assessment (N=264)

Variable	B	SE B	β
Step 1			
Grade level	1.484	0.195	0.396**
Science mark	0.053	0.008	0.353**
Gender	-0.402	0.192	-.109*
Step 2			
Grade level	1.469	0.187	.392**
Science mark	0.052	0.007	.344**
Gender	-0.295	0.186	-.080
Original vs. student	-0.637	0.130	-.285**
Student vs. expert	-0.400	0.132	-.176**

Note. $R^2=.300$ for Step 1; $R^2=.361$ for Step 2; $\Delta R^2=0.061$; * $p<.05$, ** $p<.01$

Table 8

Item Difficulty (p-values) by Item Format and Test Form

Multiple Choice				Constructed Response			
Item #	Original	Student-Modified	Expert-Modified	Item #	Original	Student-Modified	Expert-Modified
1	0.722	0.820	0.882	2	0.756	0.719	0.706
3	0.722	0.764	0.812	6	0.811	0.910	0.882
4	0.467	0.213	0.412	7	0.233	0.404	0.424
5	0.400	0.191	0.306	12	0.622	0.584	0.776
8	0.667	0.843	0.118	14	0.211	0.775	0.753
9	0.533	0.348	0.388	15	0.967	0.978	0.976
10	0.367	0.337	0.118	16	0.211	0.191	0.282
11	0.467	0.517	0.318	19	0.622	0.764	0.741
13	0.978	0.843	0.835	21	0.333	0.427	0.259
17	0.789	0.449	0.412	24	0.778	0.775	0.718
18	0.633	0.685	0.753	26	0.156	0.146	0.329
20	0.922	0.933	0.859				
22	0.556	0.584	0.588				
23	0.422	0.438	0.459				
25	0.778	0.697	0.729				
27	0.911	0.944	0.953				
28	0.822	0.775	0.788				
29	0.556	0.506	0.329				
30	0.722	0.708	0.576				
Mean	0.654	0.610	0.560	Mean	0.518	0.607	0.622
SD	0.186	0.235	0.264	SD	0.295	0.280	0.252

Table 9

Item Discrimination (point-biserial values) by Item Format and Test Form

Multiple Choice					Constructed Response				
Item #	Original	Student-Modified	Expert-Modified	Improve?	Item #	Original	Student-Modified	Expert-Modified	Improve?
1	0.218	0.180	0.019	☒☒	2	0.214	0.144	0.243	☒☑
3	0.237	0.152	0.144	☒☒	6	0.157	0.117	0.225	☒☑
4	0.328	0.121	0.370	☒☑	7	0.196	0.146	0.245	☒☑
5	0.410	0.196	0.375	☒☒	12	0.437	0.274	0.347	☒☒
8	0.316	0.312	-0.028	☒☒	14	0.297	0.164	0.279	☒☒
9	0.236	0.141	0.383	☒☑	15	0.235	0.241	0.192	☑☒
10	-0.052	0.079	0.142	☑☑	16	0.439	0.303	0.547	☒☑
11	-0.075	0.317	0.387	☑☑	19	0.185	0.220	0.223	☑☑
13	0.042	0.059	0.393	☑☑	21	0.006	0.291	0.129	☑☑
17	0.234	0.257	0.591	☑☑	24	0.179	0.202	0.316	☑☑
18	0.240	0.016	0.080	☒☒	26	0.223	-0.031	0.189	☒☒
20	0.292	0.286	0.373	☒☑					
22	0.050	-0.029	0.177	☒☑					
23	0.249	0.254	0.357	☑☑					
25	0.344	0.125	0.240	☒☒					
27	0.307	0.320	0.235	☑☒					
28	0.343	0.172	0.137	☒☒					
29	0.342	0.226	0.365	☒☑					
30	0.250	0.394	0.393	☑☑					
Mean	0.227	0.188	0.270	7 vs. 11	Mean	0.233	0.188	0.267	4 vs. 7
SD	0.137	0.113	0.159		SD	0.123	0.096	0.111	

Note: ☒☒=neither the student-modified nor the expert-modified item led to an improvement in item discrimination relative to the original, unmodified items; ☒☑=the student-modified item did not lead to an improvement in item discrimination but the expert-modified item did; ☑☒=the student-modified item led to an improvement in item discrimination but not the expert-modified item; and ☑☑=both the student-modified and expert-modified led to improvements in item discrimination; Items with point biserials close to zero or negative values suggest these items are not acting as expected with regard to the underlying construct, and should be reviewed for further modification.