University of Alberta

Using Student Verbal Reports to Evaluate Multiple Choice and Constructed Response Test Items

By

Colleen M. Heffernan

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Education

in

Measurement, Evaluation and Cognition

Department of Educational Psychology

Edmonton, Alberta

Fall 2008

Your file Votre référence
ISBN: 978-0-494-46999-6
Our file Notre référence
ISBN: 978-0-494-46999-6

ABSTRACT

The ability of novices to identify ambiguities in multiple choice and constructed response test items was investigated. Fifty-four Grade 8 and 11 students from an urban school jurisdiction were asked to identify ambiguities in test items from the School Achievement Indicators Program science assessment. Student responses were compared to the analysis of the same items by Science teachers trained in test item development. Although experts identified significantly more ambiguity, novices seemed to analyze test items in similar ways to experts. Both groups of raters identified more contextual ambiguities (the visual and semantic framework in which the item is presented to the test-taker) than structural ambiguities (the organization, formatting and testwiseness potential of the item), however multiple choice items elicited more structural concerns from both groups than did constructed response items. The findings support additional study of the evidentiary possibilities of student item analysis in the process of substantive validation.

ACKNOWLEDGEMENTS

I would like to acknowledge the dedicated support and guidance provided by the faculty of the Centre for Research in Applied Measurement and Evaluation, in particular, my supervisor, Dr. Jacqueline Leighton.

DEDICATION

This thesis is dedicated to Peggy Haggard, beloved friend and teacher, who loved unconditionally and deeply valued the rigorous pursuit of the intellect.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I - BACKGROUND

Introduction

Students sit down to write a multiple choice test. They focus on each item for one or two minutes before proceeding to the next question. During that brief time, they evaluate what the item is asking them to do and use cognitive strategies to decide on an answer. The answers that result from this complex interaction between human ability, knowledge and skill on one hand and the wording of the question on the other are used to evaluate students' performance and to infer their potential future performance. Although the interaction is between the student and the test, the conclusions and actions based on these inferences, however, have traditionally impacted the student rather than the test.

Over time, the high-stakes nature of these actions has increased, with the impact spreading across all levels of the educational system. Norway began to consider a substantial re-design of their primary school system because of their students' results on the 2003 Trends in International Mathematics and Science Study (TIMSS) and Programme for International Student Assessment (PISA) results (Gronmo et al., 2003). In 2004, with forty percent of its schools on the 'failing list' based on test scores collected under No Child Left Behind (NCLB) legislation, the city of New York restricted the number of students eligible to transfer to non-failing schools in an attempt to prevent destabilization of schools with good records (Gootman, 2004). A year earlier, the New York City Schools Chancellor threatened to remove the principals of the fifty lowest performing schools, based on a report card mark that included the results of standardized

test scores, suspension and attendance rates (Goodnough, 2003; Medina, 2003). Students have been retained in elementary school (Winerip, 2003) and denied high school diplomas across the United States based on the results of exit exams (Canedy, 2003).

Controversial test administrations, like the 2002 and 2003 Regents examinations in New York state, in which 39% and 47% of students failed physics as compared with about 10% in previous years, focus attention on the quality of the test rather than on students' performance (Gootman, 2004). The judgment of educational test quality, however, is dependent on the nature of the evidence considered for supporting claims about the knowledge and skills measured by the test.

Over the last thirty years, the nature of this evidence has greatly expanded. Test developers are increasingly expected to provide evidence for the *substance* of the test. The substance of the test refers to the alignment of a test item with the cognitive processes used by students to arrive at an answer. A test item is considered to be aligned when students use the same cognitive processes to answer the question that the item is intended to measure. If students do not use the cognitive processes that the item is designed to measure, then the item is misaligned. A test item can also be misaligned if it contains ambiguous phrasing that may result in the item being extraordinarily difficult or easy for the student for reasons unrelated to the construct, causing *construct-irrelevant variance*.

Misalignment is viewed as a serious problem for test developers and test users because it calls into question the validity of test-based inferences about student performance. Samuel Messick (1989) emphasized the importance of the substantive aspect of validity in addition to other aspects (e.g., content, consequential) in his landmark chapter on *Validity* in Robert Linn's *Educational Measurement.* More recently, Borsboom and Mellenbergh (2004), have suggested that the substantive aspect is perhaps the most important aspect of test validity. Methods for investigating the substance of a test, however, are relatively new and have yet to be evaluated. Kane (2006) proposes an argument-based validation framework in which a clear statement of the proposed interpretations and uses of a test is supported by validity evidence through a process in which the plausibility of the potential claims is evaluated. He believes it is important to move from general frameworks for analysis of validity to "providing clear guidance on how to validate specific interpretations and uses of measurement" (Kane, 2006, p.18). Validating the substance of a test requires diving into the interaction between the student and the test item to find out what students are really thinking when they answer a question. An essential part of this interaction is how students evaluate the test item itself.

The purpose of this study was to compare the identification of ambiguities in test items by novices (students) to that of experts (trained teachers) in order to establish if the results of novice 'think-alouds' add significant information to the established body of knowledge in test item development. To wit, this paper is divided into five sections. First, I will discuss the nature of the substantive aspect

of validity and the rationale for why it has become the aspect of interest. Second, I will describe cognitive methods—protocol and verbal analysis—that are considered appropriate tools for providing evidence with respect to the substantive aspect of validity. Third, I will describe the study that was done to investigate whether students provide unique information about item alignment when compared to the information provided by expert judges. Fourth, I will present the results of the study, focusing on the information provided by 54 students in Grades 8 and 11 compared to the information provided by four judges of varying expertise. Finally I will summarize and conclude with a discussion of the value of using student verbal reports to assess the alignment of test items with respect to ambiguous phrasing.

Substantive Validity

The concept of validity in educational measurement continues to evolve. At first, validity was considered to be the extent to which a test measured what it was supposed to measure (Cattell, 1946; Cureton, 1950; Kelley, 1927) and proof of validity was obtained by comparing test scores with the desired criterion. Then, Cronbach and Meehl (1955) and Messick (1989) made the case for validity being a property of the interpretation of test results, not the test itself. Because of the wide variety of possible test interpretations, this greatly expanded the concept of validity. The six aspects of validity Messick (1989) included in his definition—content, substantive, structural, external, generalizeability and consequential—gave test developers and users a large number of perspectives

to consider when preparing a validity argument. Each test interpretation is expected to have an accompanying validity argument, composed of logic and empirical evidence from all aspects of evidence relating to that interpretation. The challenge in this process is to make connections between limited samples of observation and proposed interpretations and uses (Kane, 2001, 2006).

Preparing such an argument for an educational test is challenging because of the difficulties of providing evidence to support the substantive aspect of validity. Messick (1989) envisioned the substantive aspect of validity as an extension of the qualities of relevance and representativeness that were required for content validity. For example, in the validity argument for a test of Canadian history, one would expect to find that the content of the test contains questions about important, or *relevant*, facets of history and that the test is *representative* in that it reflects the same balance of topics as the Canadian history domain. This is a complex task as experts must first agree on the proportional importance of historical topics. Professional test developers and teachers may both use curriculum documents, where available, as guides to the relevant topics and representativeness of the history domain when preparing test questions. Teachers may also use their experiences in the classroom as a guide to representativeness, emphasing topics in their tests that have been emphasized in classroom instruction. As complex as the decisions are regarding content, the substantive aspect of validity extends this complexity. The qualities of relevance and representativeness are extended to the cognitive processes of the domain in addition to the content. For example, if it is an important feature of Canadian

history to critically analyze government policy, then one would expect the test to include questions that require this analysis in the proportion that this feature exists in the domain.

The substantive aspect of validity also requires empirical evidence that the processes students use to respond to assessments are the processes that the test developer intended them to use (Messick, 1995). For every item on the Canadian history test, for example, there needs to be empirical evidence, included in the validity argument, to show that the students are responding as expected—so that the test users can be sure that students truly are recalling facts or evaluating the impact of historical events, when that is what the interpretations of students' performance are based on.

Recently, a greater focus has been given to this substantive aspect of validity, both theoretically and in a practical sense. Borsboom and Mellenbergh (2004) challenge Messick's multidimensional theory of validity:

> Validity is not complex, faceted, or dependent on nomological networks and social consequences of testing. It is a very basic concept and was correctly formulated, for instance, by Kelley (1927, p. 14) when he stated that a test is valid if it measures what it purports to measure. (p. 1061)

Borsboom and Mellenbergh (2004) see validity as a property of tests. In particular, Borsboom and Mellenbergh suggest that tests measure attributes and state that "a test is valid for measuring an attribute if and only if (a) the attribute exists and (b) variations in the attribute causally produce variations in the

outcomes of the measurement procedure" (p. 1061). Proving validity, in this instance, requires knowledge of the processes that produce item responses and how those responses are related to varied amounts of the attribute in question. Borsboom and Mellenbergh (2004) propose that knowledge of these cognitive processes can only be obtained using substantive psychological theory. Using cognitive models in test development that link item responses to stages of the development of knowledge mastery, by using the different kinds of mistakes that will be made at different stages, builds validity into the assessment instrument.

For example, a student beginning to multiply fractions, who has experience with adding fractions, might add the numerators in the following item, which would lead to the incorrect response:

2/3 X 4/5 = 8/15 (correct response)

2/3 X 4/5 = 6/15 (incorrect response based on experience adding fractions)

By creating a multiple choice item with both options included, diagnostic information can be obtained about the student who chooses the incorrect option (Leighton & Gierl, 2007). Such items also have validity evidence imbedded into them. Because the item's distracters represent incorrect processes, this design increases the likelihood that those choosing the correct option are using the process that the item intended to measure. By creating a test that includes common student misconceptions in the distracters, test developers can use

students' responses as an important aspect of the validity argument—that is, evidence that the test is measuring the skill of multiplying fractions. (Borsboom & Mellenbergh, 2004; National Research Council, 2001).

The substantive aspect of validity has become of increasing importance in the view of measurement specialists concerned with large-scale standardized assessments. In her presidential address to the National Council on Measurement in Education (NCME), Suzanne Lane (2004) outlines research that shows that curriculum standards in several states are being assessed by large scale tests at a much lower cognitive level than the objectives require. The studies Lane refers to, done by Webb in 1999 and 2002, examined the alignment between state standards and items in the corresponding state assessment in several subjects. In 1999, Webb found "that a high percentage of the four state mathematics assessments used items at a level of complexity that was below that of the corresponding objectives" (Lane, 2004, p.8). In 2002, in a study including three states, Webb found that more than half of mathematics objectives required more complex depth-of-knowledge items than the corresponding items on the assessment actually reflected.

Lane (2004) calls specifically for the increased collection of substantive evidence (for validity) in order to ensure that assessments are sufficiently cognitively challenging:

> The emphasis on using multiple-choice items to measure only basic skills sends the wrong message to the public and stakeholders and is reminiscent of former years.

> One of Messick's (1989) six aspects of construct validity is the substantive aspect that, in part, calls for studies that examine the extent to which assessment tasks are eliciting the intended cognitive processes. The accumulation of validity evidence for assessments tends to overlook this aspect. Substantive evidence should be particularly relevant under standards-based reform given that both standards and assessments should be cognitively challenging. (p. 12)

Although thinkers like Borsboom, Mellenbergh, Kane and Lane may differ on some of the definitional terms used to describe validity, they agree that the substantive aspect is of primary importance when assessing validity. Substance is also important for test developers and users who wish to understand and to measure the cognitive processes of students.

Verbal Reports - Cognitive Processing

To obtain substantive evidence of a student's problem solving process at the item level, it is necessary to find out what students are thinking as they solve the problem and arrive at an answer. Although estimates of cognitive functioning have been used in test development for some time, in the form of test specifications and models of domain mastery, these estimates can only be validated with models of task performance supported by evidence of the covert thinking processes of students (Leighton, 2004). Verbal reports, both concurrent and retrospective, are cognitive psychological methods used to collect information about how students think as they respond to problem-solving tasks (Baxter & Glaser, 1998; Ferrara, 2004; Hamilton, Nussbaum & Snow, 1997; Leighton, 2004; Norris, 1990). In concurrent reporting, students are asked to think aloud as they work on a task, creating a verbatim record of their solving of a problem (Ericsson & Simon, 1993; Taylor & Dionne, 2000; van Gog, Paas, van Merrienboer & Witte, 2005). In retrospective reporting, students are asked to remember and to relate their thought process about the solving of a task after completing it, with information retrieved from short term memory, in the case of very brief tasks, or in long term memory for lengthier tasks (Ericsson & Simon, 1993, Leighton, 2004; Taylor & Dionne, 2000; van Gog et al., 2005).

In order to ensure the validity of the information gleaned from verbal reports, Ericsson and Simon (1993) and others have developed procedures that address the criticisms regarding the validity of such reports as evidence of

cognitive processes. To wit, it was initially thought that the effect of verbalization might change the subject's thought processes. In order to minimize the effects of verbalization on the thought processes of subjects and to maximize the completeness of the record of the process, specific conditions for data collection are outlined in the literature (Ericsson & Simon, 1993; Leighton, 2004; Leighton & Gokiert, 2005; Taylor & Dionne, 2000; van Gog et al., 2005). For example, concurrent report data are collected with minimal interruptions from the interviewer (Taylor & Dionne, 2000) and retrospective report data are collected immediately following the task to ensure that the trace memory of the process is retained (van Gog et al., 2005). Ericsson (2006) reports that after the review of dozens of studies, no evidence has been found that the sequences of thoughts of participants changed when they verbalized them, as opposed to thinking silently.

It was also a concern that verbalization would result in an incomplete record of the cognitive process (Ericsson & Simon, 1993). This may occur when tasks are too easy, causing students (subjects) to utilize only automatic processing that does not leave the trace on working memory necessary for reporting. It may also occur when tasks are too difficult and overload working memory, leaving no resources for the articulation of the process (Ericsson & Simon, 1993; Leighton, 2004). Taylor and Dionne (2000) found, in their comparison of reports provided by biology and political science professors and students, that the professors, although more able to solve the problems quickly, were less able to describe their problem solving process. Taylor and Dionne explained that the professors' problem solving strategies were most likely

operating at the level of automatic processing and therefore did not leave the trace on working memory necessary for reporting. It is important that tasks chosen to elicit verbal reports of cognitive processing are of moderate difficulty for the population in question.

Even when proper methodology and tasks of moderate difficulty are used to elicit verbal reports, the nature of the information obtained from concurrent and retrospective reports would suggest that they are best used in conjunction with each other in order to ensure the validity and reliability of conclusions drawn about the problem-solving process (Taylor & Dionne, 2000; van Gog et al., 2005). Concurrent reports, which are retrieved from working memory, appear to be focused on the content or steps in the process of solving the question—the goals set at each stage, the information used in the process and the decisions made along the way (Taylor & Dionne, 2000). Retrospective reports, on the other hand, which are retrieved from long term memory, focus more on the strategies and beliefs that control the problem solving process. Taylor and Dionne (2000) state that the complementary use of concurrent and retrospective reporting provide a more comprehensive account of the problem solving process than either method used alone. Not only does each method provide information that the other does not, but the common information enables the researcher to validate the information in the concurrent report with information in the retrospective report (Taylor & Dionne, 2000). Other confirmatory methods may be used for validation of verbal reports when the task permits. van Gog et al. (2005) used eye and mouse-keyboard movements in conjunction with

retrospective reporting, in order to validate the problem solving process information received from concurrent reports with the strategy and metacognitive information from retrospective reports. Their results show some promise that this combination of techniques will increase the amount of information gleaned from verbal reports.

Using Verbal Reports for Test Development

Verbal reports of cognitive processing have wide application, having been referenced in usability studies and in the development of expert systems (van Gog et al., 2005). In educational testing research, one common use of such reports is to inform test development, in the design and re-design of test items, in order to increase the construct validity of inferences made from test results (Aikenhead, 1988; Aikenhead & Ryan, 1992; Ferrara, 2003; Ferrara, 2004; Hamilton, Nussbaum & Snow, 1997; Kane, 2006; Norris, Leighton & Phillips, 2004; Norris, 1988; Norris, 1990). Because the substantive aspect of validity is increasingly being considered important and because test developers must now provide evidence of more complex understanding than tests of basic skills and knowledge can provide, it is necessary to include the processes of more complex cognitive processes in the design phase of assessments in order to provide sufficient evidence for construct validation (National Research Council, 2001).

Verbal reports of cognitive processing are thought to provide valuable information for test development that address concerns such as the

characteristics of the dimensionality of test items, or the array of cognitive processes, knowledge and skills that an item is simultaneously testing. Hamilton Nussbaum and Snow (1997) performed a full-information item factor analysis on Grade 10 science data from the National Educational Longitudinal Study of 1988 (NELS:88) using a multidimensional item response model. This procedure confirmed three dimensions which were then tested in a verbal reporting experiment with a sample of 41 high school students of mixed socio-economic status (SES) and ethnicity, using 16 multiple choice and constructed response items from the NELS:88. Hamilton et al.'s (1997) analysis of both concurrent and retrospective verbal reports identified two important factors relating to the dimensionality of the test items. First, even when the content of the questions was similar, the underlying cognitive processes students used in their problem-solving appeared to be different for multiple choice items than for constructed response. Second, the interview data allowed for a more complete description and explanation of the dimensions obtained in the factor analysis.

Another area in which verbal reports of cognitive processing have proved valuable to test development is in illuminating the degree of common understanding between students and test developers on the purpose and meaning of test items. In the Hamilton et al. (1997) study, interviews were also done with 49 Grade 5 and 6 students who completed hands-on science tasks. The results seem to indicate that subtle differences in task instructions can have a significant effect on student performance. Hamilton et al. (1997) recommend using this kind of information to design tasks that are more likely to function in the

expected manner. Other researchers have also questioned the 'doctrine of immaculate perception' (Munby as cited in Aikenhead & Ryan, 1988), or the assumption that the student and the test developer have the same understanding of what the item means. Aikenhead and Ryan (1988) studied the responses of 27 high school students to the Views on Science-Technology-Society (VOSTS) form. Their analysis of verbal reports, obtained from the same 27 students, identified about 15 percent of ambiguous student responses as arising from the students' misunderstanding of the topic as described in the test or their inability to understand vocabulary. Leighton and Gokiert (2005) analyzed the retrospective reports of 54 students in Grades 8 and 11, responding to science items from the School Achievement Indicators Program (SAIP). They found that most students were unable to restate the items in their own words or to provide evidence of conceptual understanding in relation to the items. In other words, students were uncertain about the objective of the items, which may be important for test developers as this uncertainty may signal that items are ambiguous.

Not only does it appear that students seem to understand tasks and test items differently than test developers, there is often a mismatch of cognitive processes revealed in verbal reports, in that students use less complex processes to respond to cognitive tasks than developers had anticipated (Ferrara et al., 2003; Gierl, 1997; Glaser & Baxter, 2002; Lane, 2004; Norris et al., 2004). Researchers in the areas of critical thinking, mathematics and reading comprehension have used student verbal reports as a basis for identifying features of cognitive models used in response to test items or tasks (Norris et al.,

2004). Comparing cognitive features such as: patterns of attention, patterns of dependence on sources of information, completeness of thinking, reference to norms and principles of thinking, and metacognition to students' responses on test items and tasks have led to the conclusion that students can think poorly and yet be rewarded by a correct score and, conversely, think well and receive an incorrect score (Norris, 1988; Norris et al., 2004). Glaser and Baxter (2002) reviewed a number of standardized science assessments using verbal protocols, observation of student performance, student written work, task instructions and scoring criteria. Glaser and Baxter (2002) identified three ways in which the task design had resulted in a misalignment of cognitive processing: (a) tasks designed in a rigidly prescriptive and guided manner, so that performance was uniform across students and deep conceptual knowledge was not required, (b) tasks designed with such a high degree of openness that the goals of the task, in relation to the domain in question, were not specific enough, allowing students to use knowledge from outside the domain to complete the task, and (c) tasks designed to test only basic knowledge and skills while purporting to assess more complex concepts. In addition, when the task's design did not require the cognitive processes that it purported to assess, the scores became uninterpretable and the inferences made from them, invalid (Glaser & Baxter, 2002).

Another concern, especially with multiple choice items, is that students may be confident about their ability to answer questions that they have difficulty comprehending. Leighton and Gokiert (2005) asked 54 students a number of

retrospective questions relating to test items they had just completed. One question *(Imagine a student like yourself in your class. Do you think he or she might not understand this question? How do you know this?)* resulted in a measure of item ambiguity. Another question (*Imagine a student like yourself in your class. Do you think he or she would know how to answer this question correctly? How do you know this?*) resulted in a measure of uncertainty regarding student performance. The measure of item ambiguity did not predict uncertainty in student performance, leading to the conclusion that students may view item comprehension as independent from their ability to answer the item correctly. Where this is the case, developers may need to question the validity of inferences about student performance even when the responses are correct.

Protocol Versus Verbal Analysis

In order to provide valid and concise information from concurrent and retrospective verbal reports to inform the test item development process, rigorous procedures must be in place during the collection and analysis of data. Much of the work with verbal *think-aloud* data is in the context of protocol analysis (Newell & Simon, 1972; Ericsson & Simon, 1984; Ericsson & Simon, 1993). *Protocol analysis* focuses on the cognitive process of solving a problem or arriving at a decision. It presumes the existence of an *a priori* model and seeks to determine if the path of the problem solver matches the expected processes illustrated in the model. In contrast, *verbal analysis*, as described by Chi (1997), focuses on

the representation of knowledge that a solver possesses without an *a priori* model of what that knowledge should be.

Both protocol and verbal analysis can be used to inform the test item development process and researchers may choose between the two methods based on their confidence in an *a priori* model. Norris (1988) describes the process of the development of the Test on Appraising Observations (TAO), a critical thinking test. High school students were asked to think aloud as they attempted the items. Using an *a priori* model, the verbal data were analyzed and items were subsequently retained, modified or discarded by the developer based on the alignment of students' cognitive processing with the processing expected by the model. Another example, one which does *not* use an a priori model is the VOSTS, an attitudinal instrument designed to identify students' views on the epistemology of science. Aikenhead and Ryan (1992) were concerned that ambiguity existed in previous tests because of the potentially faulty assumption that students understood the questions and the issues in the same way as the developers. Aikenhead and Ryan (1992) describe the multi-step development process of each item on the VOSTS: (a) 50 to 70 students were asked to agree or disagree with an epistemological statement regarding science and provide a written argument to support their choice, (b) the student arguments were analyzed to identify common patterns which were in turn used to create a single multiple choice item that reflected the most common responses of students in Step A and that used the vernacular of the students, (c) a new group of students then repeated Step A as well as chose a response from the multiple choice item

developed in Step B that best reflected their views. During retrospective interviews, this second group of students was asked about the alignment between their written response and their multiple choice selection. This information was used to complete a re-development of the multiple choice item. The process of verbal analysis matched the purpose of the VOSTS, because the developers wished to establish what students' views were without an existing model of what those views should be.

Chi (1997) outlines a method for coding and analyzing verbal data without an *a priori* model, the purpose of which is to organize the data in the verbal reports so that patterns in the data related to the investigation of interest can be identified and interpreted. Depending on the purpose of the study, researchers may find it useful to randomly sample the protocols while others may wish to focus on a sub-set of the data. The grain size of the segment to be coded should be established before coding begins. Some studies will benefit from a small grain size, that is, every sentence or line of data is coded, while in other studies entire episodes of a student's response are necessary to consider in order to properly code the data. The development of a coding scheme that reflects the content and processes of the domain and allows the researcher to represent the coded results according to the appropriate set of rules, is essential to the process. Although the coding scheme should flow from what the researcher knows about the domain, it is also important to be able to adjust the coding scheme to accommodate unforeseen elements in the verbal data and thereby extend the coverage of the protocols by the coding scheme. For example, Chi

(1997) discusses coding the protocols of individuals writing computer programs. Codes might represent expected processes, like "read", "paraphrase", "compare", and "evaluate", but if the protocols indicated that individuals spent time mentally simulating the program they were writing, then "mental simulation" might be added to the list of codes. Operationalizing the codes is more or less difficult, depending on the degree to which the researcher must interpret student responses in order to assign codes. The greater degree of interpretation by the researcher increases the degree of ambiguity in the results and requires multiple raters to check the data for reliable codes to emerge. Once the data are coded, Chi (1997) recommends displaying the data patterns obtained in graphical or tabular formats with a view to interpreting the results in light of the research hypotheses. Although verbal analysis studies operate without an *a priori* model, the results of such studies may lead to the development of a cognitive processing model that may be subsequently tested in a protocol analysis study.

Verbal Reports – Metacognition

One of the important factors in a student's ability to solve problems, one for which no *a priori* model exists is metacognition. Metacognition, or the awareness, knowledge and control by an individual of his or her own thinking and learning strategies (Thomas, 2003) is important in problem solving because it allows an individual to keep track of their progress toward the goal of finding the

solution (van Gog et.al., 2005). When students identify ambiguities in test items, they are metacognitively responding to potential sources of construct irrelevant variance, by identifying obstacles to their understanding in the problem-solving path. Understanding students' ability to evaluate test items may be an important step in designing better items as well as in determining the substantive aspect of validity.

Educational researchers have frequently used self-report inventories to measure metacognition in relation to tests and environmental factors (O'Neil & Brown, 1998; Sperling, Howard, Miller & Murphy, 2003; Thomas, 2003), citing the prohibitive cost in time and money and the difficulties of scoring verbal reports. The data gathered, using self-reports, in such studies, however, are subject to social desirability.

Social desirability is a tendency of individuals to create a positive impression, manifested especially when respondents are highly motivated to achieve a goal. People may alter their responses in order to appear meritorious, either in an unconscious drive to perceive themselves in a favorable light, or in a conscious attempt to impress others (Fox & Schwartz, 2002). More over, self-reports are designed to refer in a general way to a student's performance on a test. For example, the following two questions from the Self-Assessment Questionnaire: *3. I attempted to discover the main ideas in the test questions. 15. I used multiple thinking techniques or strategies to solve the test questions.* (O'Neil & Abedi, 1996) are not applicable at the item level. Because of this, the

responses do not represent sources of information that can be used to develop or re-develop individual test items.

Researchers in cognitive psychology, however, continue to use verbal reports in conjunction with eye-movement data to gather specific information on individual tasks and processes (van Gog et. al., 2005; van Gog, Paas & van Merrienboer, 2005) that can shed light on the detailed structure of problem solving, expertise development, and the metacognitive processes that accompany this development.

Summary

The search for construct validity evidence has expanded to include the substance of the test. Researchers have used protocol and verbal analysis to examine concurrent and retrospective verbal reports from students in order to confirm cognitive processes used by students when taking test items. One of the implications is that this information should be used in test development and in test re-development. Another line of research has examined students' level of metacognition when taking tests. Both educational researchers and cognitive psychologists (Leighton & Gokiert, 2005; van Gog, 2005) have used verbal reports to identify when students are evaluating the task they are doing. Ambiguity in test items may be a source of construct irrelevant variance and something that test developers may want to consider when re-developing items.

When students identify such ambiguities, it is a metacognitive response and may reveal important sources of information for item re-development. This study will examine, using retrospective verbal reports, the degree to which students identify test item ambiguities and compare their responses to those of teachers with training in test development.

CHAPTER II - METHODS

Materials and Participants

The purpose of the study was to compare the identification of ambiguities in test items by novices (students) to that of experts (trained teachers) in order to establish if the results of novice 'think-alouds' add significant information to the established body of knowledge in test item development. In the present study, students are labeled "novices" and trained teachers are labeled "experts" only to underscore group differences in the knowledge of test item development. It is important to recognize, however, that trained teachers were not expert test developers as one would expect to find in a test development corporation since they had limited background and training. Moreover, secondary students and teachers were not classified as a novice or expert based on any criterion other than their general background experience with test development; that is, the teachers had just finished a test development course with top standing and the students had not ever been trained to develop test items during their academic training. Thus, the terms "novice" and "expert" were kept to continually remind readers that there was a difference in the background knowledge and experience between the two groups being compared in terms of their item evaluation.

The present study employed secondary data analysis—that is, student verbal reports were originally collected by Drs. Jacqueline Leighton and Rebecca

Gokiert for purposes of a previous study. The author of the present research did not participate in collecting the data. The materials used for task evaluation by both experts and novices were thirty publicly released test items from the School Achievement Indicators Program (SAIP) science assessment, a comprehensive measure of science achievement in students across Canada, administered to students in Grade 8 and Grade 11 (13-and 16-year olds) every three to five years between 1993 and 2004. Effective 2007, SAIP has been replaced by the Pan Canadian Assessment Program (PCAP). Eleven of the thirty test items were constructed response items; nineteen of the thirty test items were multiple choice items.

Novice participants.

As a reminder, for the purpose of this study, secondary students were considered to be *novices* in task evaluation because they were unschooled in formal methods of test developmentFifty-four students (14 female students and 16 male students in Grade 8; 16 female students and 8 male students in Grade 11) from four schools in an urban school jurisdiction participated in the task evaluation. Participants represented a cross section of achievement levels. Science achievement grades were available for 34 out of 54 students. The average achievement of these 34 students was 63.1 % with a standard deviation of 12.6% (range was 40.3 to 89 %). Of the remaining 20 students, 11 students were listed as honor students in a junior high school in an affluent neighbourhood and 9 students attended a suburban junior high. The thirty items were divided

into item sets containing between 4 to 6 items. Approximately 9 students were assigned to complete each of the six item sets.

Students who participated in this study were recruited from their schools with the permission of the school jurisdiction. Parental consent was obtained in writing and participants were aware that participation was voluntary and could be withdrawn at any time. All handling of human participants in this study was compliant with the University of Alberta Standards for the Protection of Human Research Participants [GFC Policy Manual, Section 66 [http://www.ualberta.ca/~unisecr/policy/sec66.html].

Expert participants.

Four teachers, three with specialized knowledge in science, participated in the study as experts. All had achieved superior performance in a required pre-service course in Educational Assessment and had worked as evaluators for subsequent administrations of the undergraduate course, judging hundreds of selected response items. As mentioned previously, for the purpose of this study, these teachers were considered experts in task evaluation because of their specialized training in test item development, although they are not professional test developers. Two of the teachers developed the item coding system (Appendix A) based on criteria for quality test items, established initial inter-rater consistency and coded the items collaboratively. Subsequently, the other two teachers used the coding system to independently rate the items.

Procedure

Sampling and segmenting student verbal reports.

Although the verbal reports contained both concurrent and retrospective responses of students to the thirty questions, only the responses to the retrospective question: *Imagine a student like yourself in your class. Do you think he or she might not understand this question? How do you know this?* were investigated in this study. The analysis consisted of three steps. At each step, every student's complete response to the question was considered a complete segment; in other words, the entirety of a student's response was considered when assigning codes. The first step was to determine which items students identified as ambiguous. Student segments were coded dichotomously, either as confusing or not confusing.

For those students who indicated the item was confusing, the second step was to identify the source of ambiguity. In step two, multiple codes were possible to reflect the possibility that students might identify multiple sources of ambiguity for each item. For example, in response to the retrospective question, a student could identify something about the item as problematic ("the pictures aren't very clear"). These segments were coded in step two as having *item ambiguity*. Students could also identify the source of the ambiguity as having to do with student characteristics and/or the classroom situation. For example, student characteristics could include ability ("some people, like their brains work at different levels"), attitude or effort ("some people don't pay

attention in class or don't put the effort in"), knowledge ("like if you don't know what cholesterol is, and you think it's a good thing, then you might get confused"), or process ("unless you like underline things to know what they're putting an emphasis on—it's sometimes hard to distinguish"). These were considered *non-item ambiguities* as were segments that identified characteristics in the classroom situation as the source of the ambiguity. For example, teachers were identified as sources of ambiguity ("depending on who their teacher was or how well they went over it in class") as was curriculum ("you just haven't been taught it yet"). In cases where students identified both item and non-item ambiguity, both sources were coded. Only segments with item ambiguity were used in step three.

The third step was to classify the student comments about item ambiguity so that they could be compared to the responses of experts in test item development. This required the development of a coding system of specific item ambiguities that could be used (a) to code all the student segments that identified item ambiguity, and (b) by the expert raters to identify elements of the item they believed to be ambiguous. The coding system (Appendix A) was developed by two of the expert raters, team A, using a number of recognized sources on constructed and selected response item development (Armstrong, 2005; Gronlund, 2005; McMillan, 2003; Norris & Ennis, 1989). The codes were organized into structural features (question and response formatting and the presence of recognized 'testwise' clues) and contextual features (the clarity or absence of visual or textual information provided by the item). Team A then

independently assessed five randomly chosen test items based on the coding scheme. The inter-rater consistency was .89 for structural features and .79 for contextual features. Following discussion, the same two raters assessed another five items, this time with an inter-rater consistency of .90 for structural and .85 for contextual features. A third trial, following additional discussion, resulted in an inter-rater consistency of .91 for structural and .93 for contextual features. Team A then coded the 30 items collaboratively.

Furthermore, Team A classified the students comments about item ambiguity using Appendix A so they could be compared to the responses of experts in test item development. Each verbal report by a student was interpreted in its entirety and assigned codes. For example, for item 13, one student said that the question was "too easy to pick the right answer because alternatives are silly". This was assigned a code of 12 (the distracters are **not** plausible and attractive to the uninformed). Another student report evaluating the same item, that also identified the responses as giving away the answer, but that additionally identified the specific wording of the stem as confusing was assigned both a code 12 and a code 18 (the stem includes **nonfunctional/irrelevant information** that may prevent the informed student from answering the item correctly). The coding of the student verbal reports is summarized in Appendix B.

A second team (B) of expert raters, two teachers with specialized training in science and test item development and evaluation then used the coding

system (Appendix A) to independently evaluate each item and to identify specific sources of ambiguity. Their ratings were compared, over 5 items with the ratings of Team A, resulting in inter-rater agreements ranging from 0.89 – 0.95 for structure and 0.83 – 0.94 for context between all four raters. Given the moderate to high levels of inter-rater agreement between raters from Team A and Team B, there was no further discussion of codes. The results section will report three expert ratings over 30 items—the first rating originating from the collaborative evaluation of Team A who developed the codes and the second and third ratings originating from Team B raters, the two teachers who independently evaluated the items. Team A raters are considered to represent a single rater because items were coded collaboratively and therefore Team A ratings can not considered independent.

CHAPTER III - RESULTS

In step one, student responses were coded dichotomously to indicate whether students found an item confusing (i.e. students identified that the item would be confusing to a student like themselves) or not confusing. The ratio of students identifying confusion was calculated for each item (see Table 1). For example, as shown in Columns 4 and 5 of Table 1, six of nine students, or 67 percent, identified confusion with item 1. The mean ratio for all thirty items was .66 (range = .22 – 1.00; SD=.20). The mean ratio, however, for the nineteen multiple choice items was .73 while the mean ratio for the eleven constructed response items was .53. Although the data were expressed as proportions by items, a Mann-Whitney U test was used, instead of a test of difference in proportions, to test the hypothesis that there was no significant difference between the groups of constructed response and multiple choice items in terms of eliciting confusion in students. The Mann-Whitney U is a nonparametric test, in which the data (in this case, proportions) are converted to ranks before the comparison is made (Glass & Hopkins, 1996). Although it is a slightly less powerful test than its parametric equivalent, the Welch t', the Mann-Whitney U is preferred in this instance because the small sample sizes in this study did not allow for strong assumptions to be made regarding the shape (normality) of the population distributions (Conover, 1980; Marascuilo & Serlin, 1988). A feature of the Mann-Whitney U test is that the Z statistic and the normal distribution provide an approximation as the sample size grows beyond 10 in either group. The

Mann-Whitney U test ($Z = -2.680$; $p < .05$) indicated that multiple choice items were found to be significantly more confusing than constructed response items.[1]

[1] The same result was confirmed using a Welch's t and a test for the difference between proportions for independent samples, where the Z-score is significant at an α of 0.05. $z = \frac{(\hat{p}_{MC} - \hat{p}_{CR}) - (p_{MC} - p_{CR})}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(.73 - .53) - (0)}{\sqrt{.66(.34)\left(\frac{1}{271} + \frac{1}{99}\right)}} = \frac{.20}{\sqrt{.2244(.0159489)}} = 3.343$

Table 1

Proportion of Students Identifying Items as Confusing and Attributing Confusion to Item Characteristics.

Item #	Item Set *	Item Type 1 = MC 2 = CR	Ratio of students identifying confusion	Ratio in percentage of students identifying confusion	Ratio of students identifying item ambiguity in confusing items	Ratio in percentage of students identifying item ambiguity in confusing items
1	1	1	6/9	0.67	1/6	0.17
2	1	2	3/9	0.33	3/3	1.00
3	1	1	8/9	0.89	4/8	0.50
4	1	1	9/9	1.00	1/9	0.11
5	2	1	7/9	0.78	6/7	0.86
6	2	2	2/9	0.22	1/2	0.50
7	2	2	7/9	0.78	2/7	0.29
8	2	1	8/9	0.89	4/8	0.50
9	2	1	6/9	0.67	4/6	0.67
10	3	1	7/9	0.78	2/7	0.29
11	3	1	7/9	0.78	4/7	0.57
12	3	2	5/9	0.56	4/5	0.80
13	3	1	5/9	0.56	5/5	1.00
14	3	2	7/9	0.78	7/7	1.00
15	4	2	4/9	0.44	1/4	0.25
16	4	2	4/9	0.44	1/4	0.25
17	4	1	5/9	0.56	2/5	0.40
18	4	1	4/9	0.44	2/4	0.50
19	4	2	5/9	0.56	1/5	0.20
20	5	1	5/8	0.63	3/5	0.60
21	5	2	6/7	0.86	6/6	1.00
22	5	1	4/8	0.50	3/4	0.75
23	5	1	7/8	0.88	2/7	0.29
24	5	2	3/8	0.38	3/3	1.00
25	6	1	9/10	0.90	2/9	0.22
26	6	2	5/10	0.50	2/5	0.40
27	6	1	6/10	0.60	1/6	0.17
28	6	1	8/10	0.80	4/8	0.50
29	6	1	7/10	0.70	2/7	0.29
30	6	1	8/9	0.89	5/8	0.63

* Note: approximately nine students responded to an item set, which included between 4 to 6 items

In step two, student responses that indicated confusion with an item, were coded to identify the source of the confusion. Because students sometimes identified more than one source of ambiguity, multiple codes were possible (see Table 2).

Table 2

Codes Identifying Source of Ambiguity in Test Items

Code	Source of Ambiguity
01	*Item*
02	Student ability
03	Student attitude or effort
04	Teacher instruction
05	Curriculum not covered
06	Student knowledge
07	Student process or strategy

Note: Codes 02 – 07 represent non-item ambiguities

A ratio was calculated for each item, where the ratio represented the number of students who identified the source of ambiguity as originating from the item itself divided by the number of students who had identified confusion in the item in step one (see Table 1). For example, although six of nine students initially identified confusion with Item 1 (See column 4 in Table 1), columns 6 and 7 of Table 1 show that only one of those six students (17 percent) specifically identified the item itself as the source of the confusion. The mean ratio for all

thirty items was .52 (range = .11 – 1.00; SD=.29), indicating that on average, half of the students who saw confusion in the items, attributed that confusion to the item itself. The mean ratio for the nineteen multiple choice items was .47 (range = .11 – 1.00; SD=.25) while the mean ratio for the eleven constructed response items was .61 (range = .20 – 1.00; SD=.35). A Mann-Whitney U test[2] indicated there was no significant difference between these two means at the .05 level (Z = -.887; p > .05). The two distributions, however, displayed different descriptive properties. The ratios of item ambiguity for the multiple choice questions were approximately normally distributed while the ratios of item ambiguity for the constructed response questions were bi-modally distributed, with one set of questions (Items 7, 15, 16 and 19) where approximately 25% of students attributed ambiguity to the item, and another set of questions (Items 2, 14, 21 and 24) in which 100% of students attributed confusion to the item (see Figure 1)[3]. Only student responses that indicated item ambiguity as the source of confusion were used in step three.

---

[2] The same result was confirmed using a Welch's t and a test for the difference between proportions for independent samples, where the Z score = -1.69 n.s.

$$Z = \frac{(\hat{p}_{MC} - \hat{p}_{CR}) - (p_{MC} - p_{CR})}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(.47 - .61) - 0}{\sqrt{.52(.48)\left(\frac{1}{125} + \frac{1}{52}\right)}} =$$

$$\frac{-.14}{\sqrt{.52(.48)(.008 + .0192307)}} = \frac{-.14}{\sqrt{.52(.48)(.0272307)}} = \frac{-.14}{.082443} = -1.69 \; n.s.$$

[3] The three expert ratings found problems with all of the items, so there are no corresponding distributions for experts because there was no variability in the proportions for experts for the 30 items. Although these histograms may be affected by the small sample size, they do seem to indicate that constructed response and multiple choice items are being evaluated differently by the student raters.

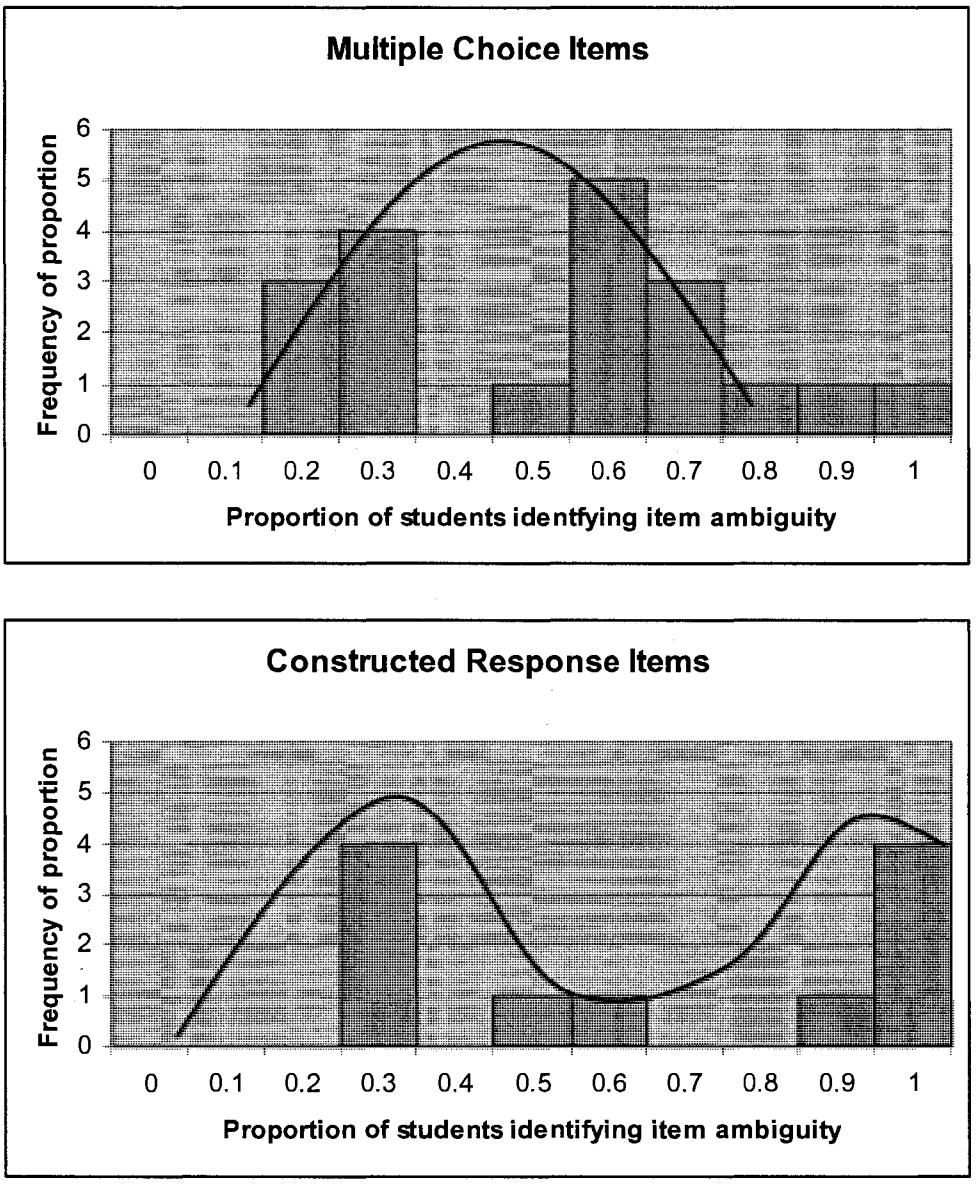

Figure 1. Proportion of Item Ambiguity Identified by Novice Raters in Verbal Reports

As mentioned previously in the Method section "*Sampling and segmenting student verbal reports*" on p. 26, in step three, a coding system (Appendix A) was developed to identify specific characteristics of ambiguity in test items. Some of these ambiguities were structural, having to do with the

organization, formatting and testwiseness potential of the item (see Table 3). Some ambiguities were contextual, having to do with the visual and semantic framework in which the item was presented to the test-taker (see Table 4). The student responses from step two that indicated item confusion originating from characteristics of the item (item ambiguity), were coded using this system. The expert raters used the same coding system to analyze the thirty items, enabling a substantive comparison between the two groups (students/novices and teachers / experts).

Table 3

Structural Ambiguity Criteria

Code	Description
	Testwiseness
01	**similar wording** exists in the stem and at least one of the responses
02	**absolutes** or **specific determiners** are used in the item (eg. Always, never, none, only)
03	there are **grammatical clues** (eg. a instead of a/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.
	Question Formatting
06	item type is **not** the best method to assess the outcome
07	the standard multiple choice layout is **not** used (eg. incomplete statement or question followed by responses)
08	there is repetitive wording in the responses that should be in the stem
09	key words (eg. best, main, negatives) are **not** emphasized by bolding, CAPS or by underlining
10	units of measurement are **not** included in the item
	Response Formatting
11	there is **not** one clearly correct answer in multiple choice responses
12	the distracters are **not** plausible and attractive to the uninformed
13	the responses are **not** homogeneous
14	the responses are **not** presented in a logical sequence eg. alpha / numeric

Table 4

Contextual Ambiguity Criteria

Code	Description
	Item Information
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
	Source Material for Interpretive Exercise
20	is **unnecessary** to correctly answer the question
21	is **lengthy** and **unreadable**
24	(Students only) – specific words identified as ambiguous

In order to compare the responses of students with the analysis of experts, a tally of codes (from Appendix A) was created for each test item (see Table 5), representing codes that were identified only by experts, codes that were identified only by novices and codes that were identified by both groups. For example, for Item 1, both novice and expert raters identified the contextual codes 18 and 20, expert raters identified the structural code 08 and novice raters contributed no unique information.

Table 5

Ambiguities Identified in Items by Expert and Novice Raters

Item	Item Set *	Item Type 1 = MC 2 = CR	Codes Identified by **Novice** raters **ONLY**	Codes Identified by **Expert** raters **ONLY**	Codes identified by **Both** Novice and Expert Raters	Number of codes identified by **Both** Novice and Expert Raters	**Total** Number of Codes assigned by all raters
1	1	1		08,	18,20	2	3
2	1	2	00, 24	17, 18, 20, 23	15	1	7
3	1	1	00, 24	13, 14 , 20	18	1	6
4	1	1	00	12, 18, 19	17	1	5
5	2	1	24	09, 18	17, 20	2	5
6	2	2		18, 20	17	1	3
7	2	2		06	18, 19	2	3
8	2	1	00, 11	08, 09, 12	01, 18	2	7
9	2	1	11	09, 18, 20, 23	12, 17	2	7
10	3	1	00	08, 09, 11, 13, 14, 17, 18		0	8
11	3	1	00, 18	13, 20		0	4
12	3	2	00, 24	20, 23	18	1	5
13	3	1	00	09, 20	12, 18	2	5
14	3	2	00, 24	09, 20	17, 18	2	6
15	4	2	24	09, 18, 20, 23		0	5
16	4	2	24	09, 18, 20		0	4
17	4	1	17, 24	11, 12, 18, 20	13	1	7
18	4	1	7	18	19	1	3
19	4	2		20	18	1	2
20	5	1	21	11, 17	20	1	4
21	5	2	00, 19	23	17	1	4
22	5	1	00, 19	13, 18, 20		0	5
23	5	1	24	08, 20	18	1	4
24	5	2	19, 24	17, 20		0	4
25	6	1	18	01, 20	13	1	4
26	6	2	00	13, 15, 20	18	1	5
27	6	1	11	12, 13, 18, 20		0	5
28	6	1	12	09, 14	18	1	4
29	6	1	00, 11	09, 12, 17, 18, 20, 23		0	8
30	6	1	11, 24	04, 09, 12, 18	15	1	7
Total			38	76		29	143

Codes listed in blue (**01 - 14**) represent structural ambiguities and codes listed in red (**15 - 24**) represent contextual ambiguities.

This organization of the data resulted in the following observations:

(a) Expert raters identified more individual ambiguities in items than novice raters (105 codes by Expert [76 + 29] : 67 codes by Novice [38 + 29]).

(b) Of the 105 coded ambiguities identified by expert raters, 64 (61%) addressed contextual concerns while 41 (39%) addressed structural concerns (see Figure 2).

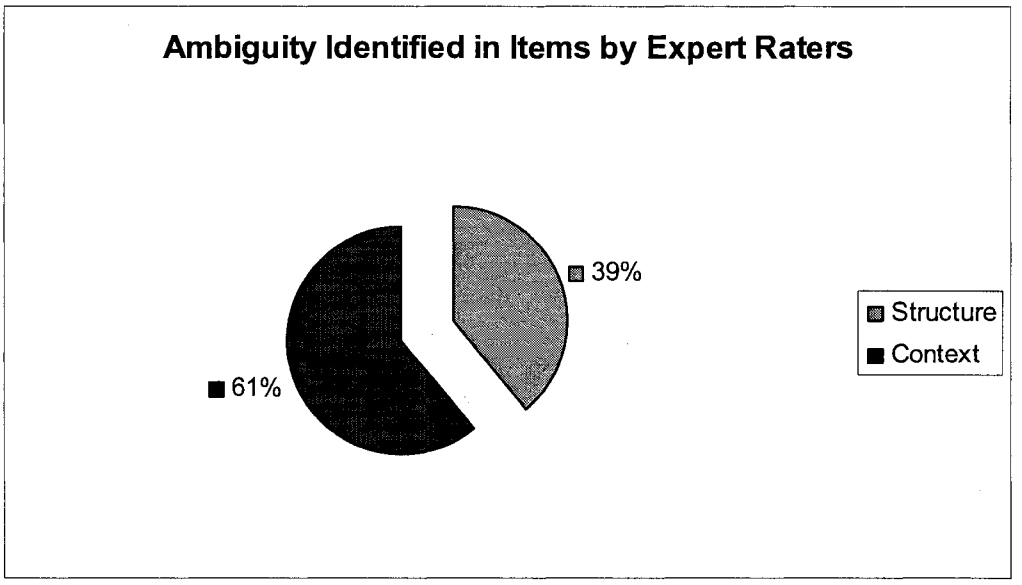

Figure 2. Ambiguity Identified in Items by Expert Raters

(c) Of the 67 concerns identified in novice responses, 13 (19%) were too general to be coded according to Appendix A. For example, statements like, "Because it doesn't tell you what kind of difference to do between them" and "It's just the answers that are a bit confusing" were coded as 00, or too general to match one of the codes. Forty-two codes (63%)

identified in the student responses addressed contextual concerns, including instances where students identified specific words as ambiguous (code 24). Twelve codes (18%) addressed structural concerns (see Figure 3).

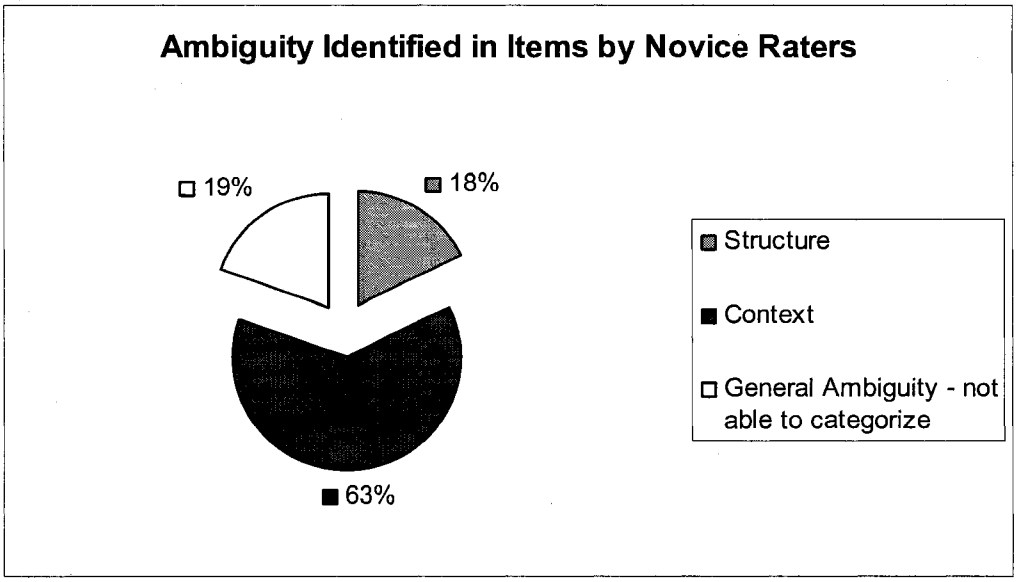

Figure 3. Ambiguity Identified in Items by Novice Raters

(d) Twenty-nine (43%) codes identified in student responses matched codes identified for the same items by expert raters. Of this group of matching codes, 24 (83%) represented contextual concerns and 5 (17%) structural concerns (see Figure 4).

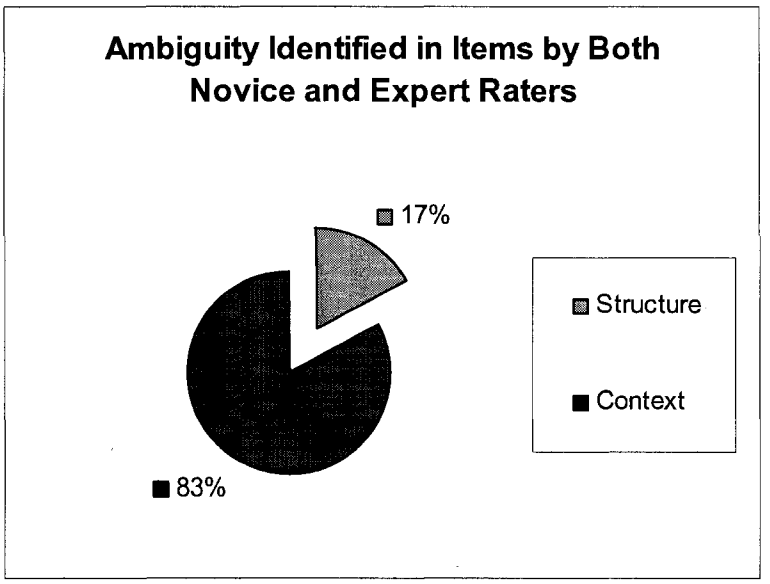

Figure 4. Ambiguity Identified in Items by Both Novice and Expert Raters

(e) When the data are reorganized to account for item type, the proportions of structural and contextual ambiguities between the two groups of raters, shift. Novice raters did not identify any structural ambiguities in constructed response items with 100% of identified ambiguities being contextual. In contrast, expert raters identified, proportionally, 14% structural : 86% contextual in constructed response items. The proportion of ambiguities identified by novice raters for multiple choice items was 33% structural: 67% contextual, while the proportion for expert raters was 49% structural : 50% contextual. Both rater groups identified about 30% more structural ambiguity in multiple choice items than in constructed response items (e.g. 33% (MC) : 0%(CR) for novice raters and 49%(MC) : 14%(CR) for expert raters, see Figure 5).

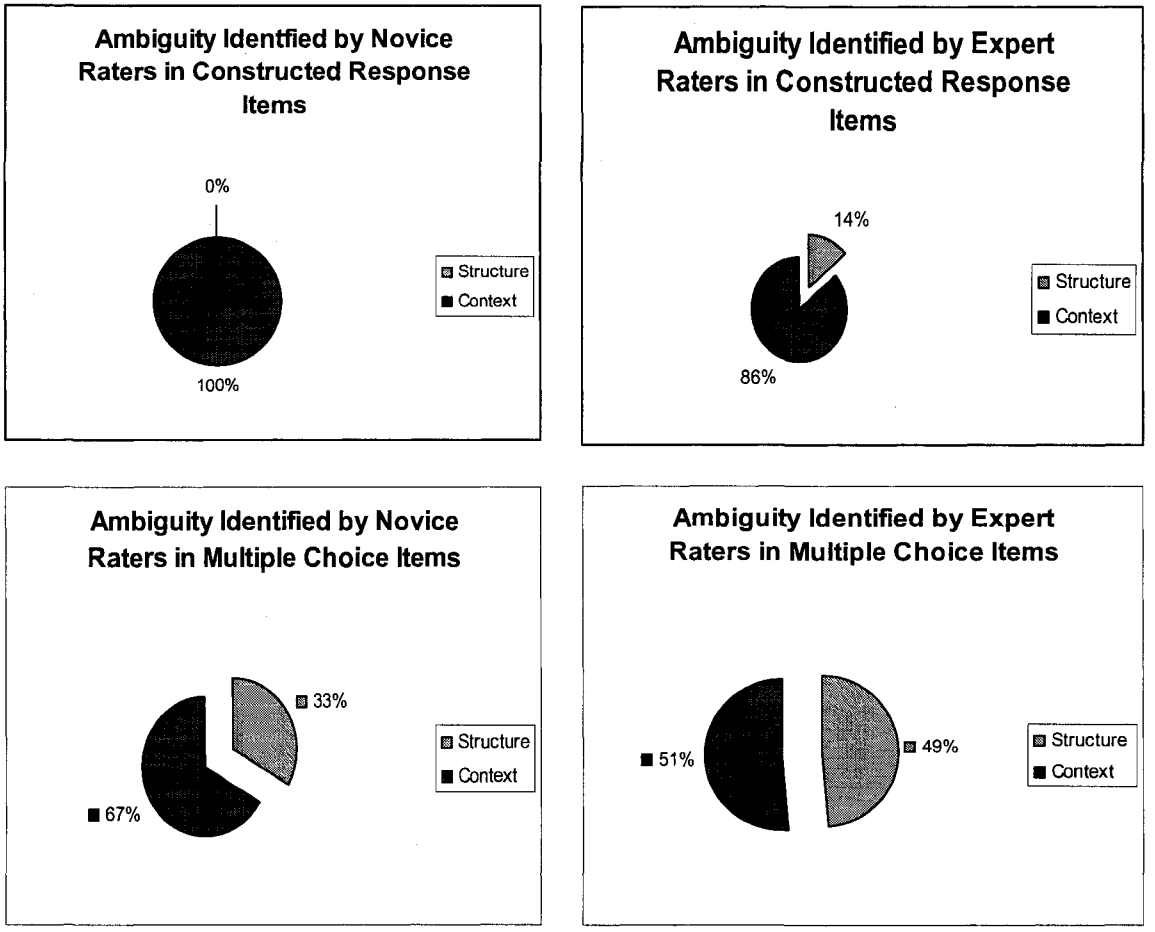

Figure 5. Ambiguities Identified by Expert and Novice Raters in Test Items

It was also of interest to determine whether there were significant differences between how experts and novices evaluated multiple choice and constructed items. Items were divided into two groups—multiple choice and constructed response. For each item, four ratings were calculated (see Table 6): (a) ratings of structural codes assigned by experts, (b) ratings of structural codes assigned by novices, (c) ratings of contextual codes assigned by experts, and (d) ratings of contextual codes assigned by novices. These ratios were based on the number of times a concern was identified by a rater divided by the number of

raters providing item ambiguity information for that item. It was important to structure the data in this manner because the number of novices providing ambiguity information varied for each item. Therefore, these ratings attempted to correct for the inequality between the number of experts and novices providing information about the items so that the respective ratings could be compared. For example, for item 1, a multiple choice item: (a) all three expert raters[4] each identified one structural code (08 – repetitive wording in the responses), divided by the number of raters providing information, resulting in a rating of 3/3, or 1.00, (b) no novice raters identified structural codes, divided by one novice providing information on this item, resulting in a rating of 0/1, or 0.00, (c) all three expert raters each identified two contextual codes (18 – unnecessary wording in item stem, 20 – unnecessary source material), divided by the number of experts, resulting in a rating of 6/3, or 2.00, (d) one novice rater identified two contextual codes (18, 20), divided by one student providing information on this item, resulting in a rating of 2/1, or 2.00.

The following tests were performed on the data as organized in Table 6. First, in order to test if there was a difference between raters in their evaluations of items for structure and context, a Wilcoxon Signed Ranks test was conducted. A Wilcoxon Sign Ranks test was conducted because it was necessary to consider that the ratings were paired, that is, each item had a structure rating from experts and a structure rating from novices (as well as a context rating from

[4] The data from 3 raters were used—Rater 1 represented the collaborative ratings of the 30 items by the two experts in Team A. Raters 2 and 3 were the independent ratings by the two experts in Team B.

experts and a context rating from novices). The Wilcoxon Signed Ranks test is a nonparametric test for paired data that tests if the medians of the two pairs come from the same population (Conover, 1980). It is preferred, in this case, to the paired *t test* because of the small sample size (Glass & Hopkins, 1996). A feature of the Wilcoxon Signed Ranks test is that the Z statistic and the normal distribution provide an approximation as the sample size grows beyond 10 in either group. One Wilcoxon test (using Columns 2 and 3 in Table 6) indicated that experts provided significantly more structural information than novices ($Z = -4.120$; $p < .05$). Another Wilcoxon test (using Columns 4 and 5 in Table 6) indicated that experts also provided significantly more contextual information than novices ($Z = -4.608$; $p < .05$),

Second, in order to determine if there was a difference within rater in the proportion of contextual and structural ratio of codes provided by each group, a second pair of Wilcoxon tests was conducted. The first test (using Columns 3 and 5 in Table 6) indicated that novices provided significantly more contextual information than they did structural information ($Z = -3.793$; $p < .05$). The second test (using Columns 2 and 4 in Table 6) indicated that experts also provided significantly more contextual information than they did structural information ($Z = -3.723$; $p < .05$).

Third, in order to determine if there were differences between multiple choice and constructed response items in structure and contextual codes, four Mann-Whitney U tests (using Columns 2, 3, 4, and 5 of Table 6) were conducted.

The Mann-Whitney U tests the differences of the mean rankings between two independent groups (multiple choice and constructed response) and is more appropriate here than the Welch t' because of the small sample size (Conover, 1980, Marascuilo & Serlin, 1988). The first test (Column 2 of Table 6) indicated that multiple choice items elicited more structural concerns from experts than constructed response items ($Z = -3.208$; $p < .05$). The second test (Column 3 of Table 6) indicated that that multiple choice items elicited more structural concerns from novices than constructed response items ($Z = -2.830$; $p < .05$). There was no significant difference, however, between multiple choice and constructed response in the amount of contextual concerns elicited from experts ($Z = -.998$; $p > .05$). Neither was there a difference between multiple choice and constructed response in the amount of contextual concerns elicited from novices ($Z = -1.285$; $p > .05$).

Table 6

Ambiguity Ratings of Test Items by Novice and Expert Raters

Multiple Choice Items				
	Expert - Structure	**Novice - Structure**	**Expert – Context**	**Novice - Context**
Item #	Ratings: # of codes / #of raters	Ratings: # of codes / #of raters	Ratings: # of codes / #of raters	Ratings: # of codes / #of raters
1	1.00	0.00	2.00	2.00
3	0.67	0.00	2.00	0.50
4	0.33	0.00	2.33	1.00
5	1.00	0.00	2.33	1.00
8	3.00	0.50	2.67	0.50
9	1.33	0.75	2.33	0.50
10	4.00	0.00	2.00	0.50
11	1.67	0.00	1.33	0.75
13	1.33	0.60	2.00	0.60
17	2.33	0.50	2.00	1.00
18	0.00	0.50	1.33	0.50
20	1.00	0.00	1.33	1.33
22	0.67	0.00	2.00	0.67
23	0.33	0.00	2.00	1.00
25	0.67	0.50	1.00	0.50
27	2.67	1.00	2.33	0.00
28	0.67	0.25	2.00	0.75
29	1.67	0.50	2.33	0.00
30	2.67	0.40	2.33	0.60
Constructed Response				
Item #				
2	0.00	0.00	4.00	0.67
6	0.00	0.00	2.33	1.00
7	0.67	0.00	2.33	1.00
12	0.00	0.00	2.00	0.75
14	1.33	0.00	2.67	0.71
15	0.67	0.00	2.00	1.00
16	0.67	0.00	2.00	1.00
19	0.00	0.00	2.00	1.00
21	0.00	0.00	0.33	0.50
24	0.00	0.00	2.00	1.00
26	0.33	0.00	3.00	0.50

CHAPTER IV - DISCUSSION

The objective of this study was to compare the identification of ambiguities in test items by novices (students) to that of experts (trained teachers) in order to establish the extent of task evaluation in novices to determine if the results of novice 'think-alouds' add significant information to the established body of knowledge in test item development. Ambiguities are potential obstacles to the valid assessment of students' knowledge and skills. The extent to which students can provide information about these obstacles to their own performance is an important factor for educators and test developers to consider when designing better test items. The evaluation of test items is a metacognitive process and it was hoped this process would be revealed by comparing student verbal reports with the analysis of teachers trained in test item development. This discussion will compare student results to those of the teacher/experts in order to reveal the type and quality of information students can provide about test item ambiguity.

The student verbal reports indicated that students experienced considerable confusion when approaching these science items—on average, 66 percent of the 270 verbal reports (9 students X 30 items or 54 students X 5 items) indicated that students thought that 'a student like them' would find the item confusing. Only about half of this confusion, however, appeared, in the minds of students, to be directly related to the test item itself (mean ratio of .52 over 30 items, see p.31). The other half of the confusion, students attributed to

internal sources—students didn't know enough, they used the wrong strategy, had poor attitudes or reduced ability. Although this was not the focus of this study, further research may be necessary to explore students' attribution of confusion with test items to internal characteristics as a factor in academic self-efficacy or in mind-set.

Students found multiple choice items significantly more confusing than constructed response items (.73 versus .53, see p. 28), therefore it might be expected that students would also attribute more item ambiguity to multiple choice items. However, this was not the case and there was no significant difference in the attribution of ambiguity to the item between the two question types (.47 versus .61, see p.31). One possible explanation for this is that students may view the multiple choice format as more objective and therefore be less likely to attribute confusion to the item. Another possible explanation relates to the types of ambiguity present in multiple choice items. Ambiguities identified by expert raters in multiple choice items were equally divided into contextual and structural concerns, while only a third of ambiguities identified by novices for multiple choice items were structural in nature (see p. 40). The codes for structural ambiguity were based on guidelines for writing test items that provide clarity to the test taker and eliminate the use of testwiseness to answer the question (Armstrong, 2005; Gronlund, 2005; McMillan, 2003; Norris & Ennis, 1989). As these guidelines are well known to experts, but not necessarily to novices, it is therefore not surprising that experts would identify proportionally more ambiguities according to structural criteria since students may not have

known how to articulate problems with the structure of the multiple choice item type. The contextual criteria, although also based on accepted test development principles, are more general and do not require prior knowledge in test development to articulate so it may have been easier for students to communicate their ideas about contextual problems with the item.

The study showed that teachers identified more structural and contextual ambiguities in the test items than students. While this is understandable, as a result of teachers' expertise in test item development, it also speaks to the efficacy of using student verbal reports as a primary source of information for test item revision. The acquisition and analysis of student verbal reports is a time consuming process that may not be worth investing in unless there is evidence that it will yield results. Although teachers identified more ambiguities than students did across item types, the analysis seems to indicate that students were able to provide important information about test items. Forty three percent of the codes in the student reports matched codes that the teachers had identified for those items (29 / 67 student codes, see Table 5). This suggests that students are able to engage in the metacognitive process of evaluating test items in a similar way to trained teachers. There is also evidence that they can provide new information. Of the codes in the student data, 38 % represented contextual or structural ambiguities not identified by experts. Half of this new information, however, represented specific words that students identified as ambiguous (see p.37). This 'word' ambiguity may be associated with reading comprehension, however, and may not indicate test item construction errors. It was not within the

scope of this study to determine the usefulness of this new information provided by students, but further study in this area would be recommended in order to discover the breadth and depth of what students can tell us about the interactions between themselves and test items.

Constructed response and multiple choice items seem to have different ambiguity profiles. Both novices and experts found more structural concerns in multiple choice items than in constructed response questions, even though, over all the items, both groups of raters (experts and novices) identified more contextual concerns than structural. This is perhaps due to the nature of the structural codes (Appendix A) in that some of them relate specifically to the quality of distracters in selected response items and would therefore not be applicable to constructed response questions.

These results take a small step into uncharted waters—it cannot be determined from this study alone how much information students can provide about the assessment tools used to evaluate their performance or what other cognitive processes students may be able to articulate. However, that students were able to identify ambiguities in similar ways to experts does indicate that further research in this area may be valuable for test developers and for classroom educators. One step would be to empirically test the value of the information provided by students by using these data to modify test items and assess subsequent performance. One possible limitation of this study was that because the students were asked an open-ended question, they really didn't

know what the interviewer was looking for (*Imagine a student like yourself in your class. Do you think he or she might not understand this question? How do you know this?*). Even so, some of the students were able to match the concerns of expert raters in identifying ambiguities. The question remains as to how much information students could provide if the criteria was given to them in advance and they were trained in the characteristics of quality test items.

From the point of view of test developers, the results of this study seem to indicate that soliciting verbal reports from students may be redundant, because expert raters provide significantly more information about structural and contextual ambiguities and because the evidence seems to suggest that experts and novices evaluate the items in similar ways. It should be noted, however, that the expert raters in this study were also secondary school teachers, actively working with students and perhaps more intuitively aware about how they might comprehend test items. This specialized knowledge of students may account for the overlap of codes between experts and novices. In addition, the large amount of ambiguities identified by both groups of raters may also indicate that items which are developed without the knowledge of how students think, either from student reports or 'student-savvy' experts, may not be able to provide strong evidence that the inferences based on the results are valid.

As well, the test development community is being asked to provide more evidence for the validity of tests, especially the substantive aspect which confirms the cognitive processes students must use to correctly answer test

items (Borsboom, 2006; Kane, 2006; Lane, 2004). In this study, students were able to articulate ambiguities in test items in similar ways to expert raters. Although it was not the main focus of the study, students identified, in their attributions of ambiguity, problem solving strategies used to analyze and answer test items. For example, the following student identifies 'overthinking' as a process that students might use that results in reduced performance on test items:

> I: Imagine a student like yourself in your class. Do you think he or she might not understand this question?
>
> S: Mmm, possibly if the student kind of over thought it—
>
> I: Okay.
>
> S: —and thought that when they said "smallest particles," it meant like atoms or something like that.
>
> I: And how do you know this?
>
> S: Um, just because that if I over thought the question—
>
> I: Mm-hmm.
>
> S: —if I thought that it was—if I—the way they—if they worded it kind of so I thought it was a trick question—
>
> I: Mm-hmm.
>
> S: —then if I over thought like that, I'd probably answer—I might answer it differently.

The information articulated by students about their approaches to problem solving suggest that further studies may be devised which would allow students verbal reports to become evidence of the substantive aspect of validity.

For classroom educators, the results of this study may have a two-fold impact. While test developers are concerned primarily with the reliability and validity associated with their tests, classroom teachers are not only concerned with the ability of their assessments to provide valid inferences about their students' performance, they are also concerned with increasing their students' ability to think evaluatively about their own learning. The lack of teacher self-efficacy and knowledge in regards to test development (Airasian, 1991; Stiggins, 1986) often leaves teachers reusing poorly written items without the knowledge or process to improve them. An evaluative process that recognizes students' ability to identify problem elements in test items may lead to improved test items, improved test-based inferences, increased teacher knowledge and confidence in item development, and increased student metacognition.

REFERENCES

Aikenhead, G. S. (1988). An analysis of four ways of assessing student beliefs about STS topics. *Journal of Research in Science Teaching, 25*(8), 607-629.

Aikenhead, G. S., & Ryan, A. G. (1992). The development of a new instrument: "Views on science-technology-society" (VOSTS). *Science Education, 76*(5), 477-491.

Airasian, P. W. (1991). Perspectives on measurement instruction. *Educational Measurement: Issues and Practice, 10(5)*, 13-16.

Baxter, G. P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice, 17*, 37-45.

Borsboom, D. (2006). Can we bring about a velvet revolution in psychological measurement? A rejoinder to commentaries. *Psychometrika*, 71(3), 463-467.

Borsboom, D., & Mellenbergh, G. J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061-1071.

Canedy, D. (2003, May 13, 2003). Critics of graduation exam threaten boycott in florida. *New York Times,*

Cattell, R.B. (1946). *Description and measurement of personality*. New York: World Book Company.

Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences, 6*(3), 271-315.

Conover, W.J. (1980). *Practical nonparametric statistics*. Toronto: John Wiley and Sons.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Cureton, E.E. (1950). Validity. In E.F. Lingquist (Ed.), *Educational Measurement.* Washington, DC: American Council on Education.

Ericsson, K. A. (2006). Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks. In K. A. Ericsson (Ed.), *The Cambridge handbook of expertise and expert performance* (pp. 223-241). New York: Cambridge University Press.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis*. Cambridge, MA: The MIT Press.

Ericsson, K. A., & Smith, J. (2002). Prospects and limits of the empirical study of expertise: An introduction. In D. J. Levitin (Ed.), *Foundations of cognitive psychology: Core readings* (pp. 517-550). Cambridge, Mass.: MIT Press.

Ferrara, S., Duncan, T., Freed, R., Velez-Paschke, A., McGivern, J., & Mushlin, S., et al. (2004). Examining test score validity by examining item construct validity: Preliminary analysis of evidence of the alignment of targeted and observed content, skills, and cognitive processes in a middle school science assessment. *Annual Meeting of the American Educational Research Association,* San Diego, CA.

Ferrara, S., Duncan, T., Perie, M., Freed, R., McGivern, J., & Chilukuri, R. (2003). Item construct validity: Early results from a study of the relationship between intended and actual cognitive demands in a middle school science assessment. *Annual Meeting of the American Educational Research Association,* Chicago, IL.

Fox, S., & Schwartz, D. (2002). Social desirability and controllability in computerized and paper-and-pencil personality questionnaires. *Computers in Human Behavior, 18*(4), 389-410.

Glaser, R., & Baxter, G. P. (2002). Cognition and construct validity: Evidence for the nature of cognitive performance in assessment situations. In H. I. Braun, D. N. Jackson & D. E. Wiley (Eds.), *The role of constructs in psychological*

and educational measurement (pp. 179-192). Mahwah, NJ: Lawrence Erlbaum Associates.

Glass, G.V. & Hopkins, K.D. (1996). *Statistical methods in education and psychology.* Toronto: Allyn and Bacon.

Goodnough, A. (2003, May 23). Trainer of school principals says he almost quit, twice. *New York Times.* Retrieved May 24, 2003, from http://www.nytimes.com/

Gootman, E. (2004, July 17). New york city will limit chance to leave failing schools. *New York Times.* Retrieved July 20, 2004 from http://www.nytimes.com/

Gootman, E. (2004, January 10). Thousands pass regents tests under revised scoring. *New York Times.* Retrieved January 13, 2004 from http://www.nytimes.com/

Gronlund, N. E. (2005). *Assessment of student achievement* (7th ed.). Boston, MA: Pearson Custom Publishing.

Grønmo, L. S., Bergem, O. K., Kjærnsli, M., Lie, S., Turmo, A., & Olsen, R. V., (2003). *Norwegian reports from TIMSS and PISA 2003: Short english versions.* Oslo, Norway: Institute for Teacher Education and School Development, University of Oslo.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CF: Sage Publications.

Hamilton, L. S., Nussbaum, E. M., & Snow, R. E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education, 10*(2), 181-200.

Kane, M. T. (2006). 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.

Kane, M.T. (2006). In praise of pluralism. A comment on Borsboom. *Psychometrika*, 71(3), 441-445.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319-342.

Kelley, T.L. (1927). *Interpretation of educational measurements*. New York: Macmillan

Lane, S. (2004). Validity of high-stakes assessment: Are students engaged in complex thinking? *Educational Measurement, Issues and Practice, 23*(3), 6-14.

Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice, 23*(1), 6-15.

Leighton, J.P. & Gierl, M.J. (2007). Why cognitive diagnostic assessment? In J.P. Leighton & M.J. Gierl (Eds.), *Cognitive diagnostic assessment for education : Theory and applications* (pp. 3 - 18). New York: Cambridge University Press.

Leighton, J. P., & Gokiert, R. J. (2005). The cognitive effects of test item features: Informing item generation by identifying construct irrelevant variance. *Annual Meeting of the National Council on Measurement in Education (NCME),* Montreal, Quebec, Canada.

Marascuilo, L.A. & Serlin, R.C. (1988). *Statistical methods for the social and behavioral sciences.* New York: W.H. Freeman and Company.

McMillan, J. (2001). *Classroom assessment: Principles and practice for effective instruction*. Boston, MA: Allyn and Bacon.

Medina, J. (2003, January 2, 2003). Chancellor gives out report cards, failing 50 principals. *New York Times,*

Messick, S. (1995). Validation of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meanings. *American Psychologist, 50*, 741-749.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Washington, DC: The American Council on Education and the National Council on Measurement in Education.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*, 5-12.

National Research Council. (2001). In Pellegrino J. W., Chudowsky N. and Glaser R. (Eds.), *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Norris, S. P. (1995). Format effects on critical thinking test performance. *The Alberta Journal of Educational Research, 41*(4), 378-406.

Norris, S. P. (1990). Effect of eliciting verbal reports of thinking on critical thinking test performance. *Journal of Educational Measurement, 27*(1), 41-58.

Norris, S. P. (1988). Controlling for background beliefs when developing mulitiple-choice critical thinking tests. *Educational Measurement, 7*(3), 5-11.

Norris, S. P., & Ennis, R. H. (1989). Making your own multiple-choice critical thinking tests. *Evaluating critical thinking* (pp. 101-158). Pacific Grove, CA: Midwest Publications.

Norris, S. P., Leighton, J. P., & Phillips, L. M. (2004). What is at stake in knowing the content and capabilities of children's minds? A case for basing high stakes tests on cognitive models. *Theory and Research in Education, 2*, 283-308.

O'Neil, H. F., & Abedi, J. (1996). Reliability and validity of a state metacognitive inventory: Potential for alternative assessment. *ournal of Educational Research, 89*(4), 234-245.

O'Neil, H. F., & Brown, R. S. (1998). Differential effects of question formats in math assessment on metacognition and affect. *Applied Measurement in Education, 11*(4), 331-351.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. *Review of Research in Education, 24*, 307-353.

Sperling, R. A., Howard, B. C., Miller, L. A., & Murphy, C. (2002). Measures of children's knowledge and regulation of cognition. *Contemporary Educational Psychology, 27*, 51-79.

Stiggins, R. J., Conklin, N. F., & Bridgeford, N. J. (1986). Classroom assessment: A key to effective education. *Educational Measurement: Issues and Practice, 5(2)*, 5-17.

Taylor, K. L., & Dionne, J. (2000). Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology, 92*(3), 413-425.

Thomas, G. P. (2003). Conceptualisation, development and validation of an instrument for investigating the metacognitive orientation of science classroom learning environments: The metacognitive orientation learning environment scale - science (MOLES-S). *Learning Environments Research, 6*(2), 175-197.

van Gog, T., Paas, F., & van Merrienboer, J. J. G. (2005). Uncovering expertise-related differences in troubleshooting performance: Combining eye movement and concurrent verbal protocol data. *Applied Cognitive Psychology, 19*(2), 205-221.

van Gog, T., Paas, F., van Merrienboer, J. J. G., & Witte, P. (2005). Uncovering the problem-solving process: Cued retrospective reporting versus concurrent and retrospective reporting. *Journal of Experimental Psychology, 11*(4), 237-244.

Winerip, M. (2003, May 21). A pupil held back, A heavier burden. *New York Times.* Retrieved May 24, 2004 from http://www.nytimes.com/

APPENDICES

Appendix A - Coding System for Item Ambiguities

Appendix B - Coding of Student Verbal Reports

Item	Structure checklist	Context checklist
	Testwiseness __ **similar wording** exists in the stem and at least one of the responses __ **absolutes** or **specific determiners** are used in the item (eg. Always, never, none, only) __ there are **grammatical clues** (eg. a instead of a/an) in the stem and/or responses __ the **length and/or detail** of the correct answer is significantly **greater** than the distracters __ the correct answer is stated in **'textbook'** or **stereotyped language** that enables the uniformed student to select it. **Question Formatting** __ item type is **not** the best method to assess the outcome __ the standard multiple choice layout is **not** used (eg. incomplete statement or question followed by responses) __ there is repetitive wording in the responses that should be in the stem __ key words (eg. best, main, negatives) are **not** emphasized by bolding, CAPS or by underlining __ units of measurement are **not** included in the item **Response Formatting** __ there is **not** one clearly correct answer in multiple choice responses __ the distracters are **not** plausible and attractive to the uninformed __ the responses are **not** homogeneous __ the responses are **not** presented in a logical sequence eg. alpha / numeric	**Item Information** __ the item contains **visuals** that are **unnecessary** or **unclear** __ the item tests **multiple** concepts, skills or problems __ the stem **lacks necessary information** that is required for an informed student to answer the item correctly __ the stem includes **nonfunctional/irrelevant information** that may prevent the informed student from answering the item correctly __ the item makes assumptions about prior knowledge or uses language that introduces bias toward a specific group **Source Material for Interpretive Exercise** __ is **unnecessary** to correctly answer the question __ is **lengthy** and **unreadable** __ is **irrelevant** to the learner outcome **Other:**_____ _____ _____ **Taxonomy Mismatch (circle identified level)** Level of Question 1 – knowledge 2 – comprehension 3 – application 4 – analysis 5 – synthesis 6 – evaluation Level of Outcome 1 – knowledge 2 – comprehension 3 – application 4 – analysis 5 – synthesis 6 – evaluation **Content Mismatch** __ the item is testing material that is **not** explicitly connected to the identified learner outcome **Highly ambiguous Words** non-scientific words that have more than one connotation **OR** that a secondary school student would not understand: _____ _____ _____

Appendix B - Coding of Student Verbal Reports

Item	Student ID	Identified Ambiguities	Code
1	• 046-11	• there is unspecified "irrelevant information that makes the question tricky"	18 20
2	• 005-11	• the word "between"	24
	• 042-11	• "it doesn't tell you what KIND of difference to do between them"	00
	• 046-11	• pictures are NOT clear	15
3	• 005-11	• the wording "over the course of the night" is confusing because it seemed like you had to BE there to know the answer	18
	• 049-11	• answers are a bit confusing	00
	• 019-08	• there's a lot of possible answers	00
	• 040-08	• the word "constellation"	24
4	• 046-11	• "doesn't describe what you're trying to figure out"	17
	• several	• haven't covered this material yet	00

Item	Student ID	Identified Ambiguities	Code
5	• 044-11	• the word 'composed	24
	• 052-11	• picture doesn't show anything AND irrelevant info: "Patrick notices that...pond" AND students will think there's a hidden trick	20
	• 016-08	• tricky question because just because sand is smaller, doesn't mean it has the smaller particles	17
	• 021-08	• trick question – students might 'overthink' eg. Atoms	17
	• 037-08	• confusion between clay and sand	17
	• 003-11	• the materials are confusing because of the way people think about solids	17
6	• 051-11	• Isn't clear whether it starts small and goes up or starts large and goes down the food chain	17
7	• 052-11	• Irrelevant info about the pond could cause confusion	18
	• 038-08	• "soft drink" is more American; in Canada, we say "pop"	19

Item	Student ID	Identified Ambiguities	Code
8	• 003-11	• "it's a very confusing kind of statement. They have a lot of things that are thrown in there to throw you off." No specifics	00
	• 052-11	• word 'analysis' in stem and in 2 of the responses AND they also use 'analysis', then 'results' –confusing because they are using 2 different words for the same thing	01 18
	• 016-08	• trick question because they use 'similar/different' – then you have to think of 'method' and 'analysis'	18
	• 021-08	• A and C are both plausible answers	11
9	• 003-11	• "They have answers like 'repeat their measurements' which would seem right to somebody who didn't know the thing about averaging"	11
	• 051-11	• Confusion between 'repeating' and 'averaging' because both are used	11
	• 044-11	• Seems like there is missing info eg. Time of day AND able to eliminate 2 distracters right away – no specifics	17 12
	• 016-08	• The large difference in temperature is confusing – it makes the student want more background info eg. Heat in the room	17

Item	Student ID	Identified Ambiguities	Code
10	• 053-11	• ultraviolet light would be attractive to uninformed	00
	• 035-08	• 'major advantage' could cause confusion	00
11	• 035-08	• 'since light can be polarized' does NOT relate to the answer of the question	18
	• 012-08	• 'since light can be polarized' – is confusing wording	18
	• 008-11	• 'since light can be polarized' is confusing, causing student to question prior knowledge AND previous question (10) caused confusion	18
	• 022-08	• wording is confusing, eg. "what types of waves are light waves?"	00
12	• 047-11	• "spoiling" is a source of comprehension confusion	24
	• 012-08	• the word 'spoiling' could cause confusion	24
	• 053-11	• difficult to describe a way to slow down spoiling of food	00
	• 035-08	• 'back story' seems unnecessary and could cause confusion	18

Item	Student ID	Identified Ambiguities	Code
13	• 047-11	• question is written in a 'confusing' manner AND responses help to give the answer AND the word 'explain' causes confusion—it should be 'restate'	00 12 18
	• 053-11	• wording of question is confusing AND distracters are misleading, specifically B	00 12
	• 035-08	• 'explain' and the question do NOT match—the two different conflicting parts to the question should be joined	18
	• 022-08	• description doesn't match question, specifically teacher 'explains', but they are just asked for a fact, with NO explanation	18
	• 008-11	• "too easy to pick the right answer because alternatives are silly"	12
14	• 045-11	• "question is poorly written and could be read incorrectly" – no specifics	00
	• 047-11	• too wordy for some students to follow	00
	• 053-11	• "too many answers—don't know what to concentrate on"	00
	• 034-08	• overall reading level is confusing AND the word 'composition' is confusing	18 24
	• 035-08	• word 'composition' is	24

Item	Student ID	Identified Ambiguities	Code
		confusing AND the back story side tracks people AND the question is wordy	18
	• 036-08	• not enough info to answer question correctly, i.e. need to know the type of rock and acid AND question is TOO broad	17
	• 008-11	• 'entire question' is confusing, eg. the description of what is reacting with acid is NOT clear	00
15	• 031-08	• confusing words: organisms, aquatic and unaided	24
16	• 031-08	• words hard to understand: sexually and asexual	24
17	• 033-08	• the question doesn't give enough information i.e. it does NOT state the environmental changes AND it needs a specific example to make the question clearer	17
	• 031-08	• the responses are misleading (no specifics) AND the word 'species' is too broad	13 24
18	• 031-08	• question position is in a strange place	07
	• 033-08	• students might lack experience with rearview mirrors	19

Item	Student ID	Identified Ambiguities	Code
19	• 033-08	• confusing because the question uses minutes and the answer uses hours	18
20	• 010-11	• way too much irrelevant information i.e. '2 hrs time' AND don't need the long back story	20 21
	• 028-08	• story is TOO long	21
	• 030-08	• too much reading of useless information causing boredom and loss of focus	20
21	• 007-11	• "both teams carry out temp, soil, water and pH tests" is confusing, specifically the temperature part	00
	• 010-11	• reading difficulty could cause confusion	19
	• 043-11	• there is NOT enough information about each team's results, causing confusion	17
	• 013-08	• "the two different teams with the same procedures getting different results is confusing"	00
	• 029-08	• reading comprehension	19
	• 030-08	• not enough information to make a single clear conclusion—too many possible answers	00

Item	Student ID	Identified Ambiguities	Code
22	• 010-11	• "the alternatives do NOT have anything to do with the question" AND grammar gives it away	00
	• 028-08	• reading comprehension could cause issues	19
	• 029-08	• reading level way too high (no specifics)	19
23	• 007-11	• 'plate tectonics' is confusing	24
	• 010-11	• the link to the other story is silly and has nothing to do with the question	18
24	• 010-11	• wording is confusing – 'nature of the accident'	24
	• 013-08	• reading level is very difficult and could cause confusion	19
	• 028-08	• reading level may cause confusion	19
25	• 006-11	• responses are similar "close to each other"	13
	• 017-08	• extraneous information in question, specifically a) 200 litres and b) metal barrel	18
26	• 006-11	• options are confusing, specifically a) plastic and b) cardboard	00
	• 017-08	• the word "hardest" has two meanings AND toxic waste is NOT recyclable	18
27	• 006-11	• responses B and D may both be correct	11

Item	Student ID	Identified Ambiguities	Code
28	• 001-11	• distracting information about power lines	18
	• 006-11	• response C doesn't specify the location of the tectonic plates	12
	• 017-08	• extra information that you don't need to know, specifically "Bay of Fundy tides are the highest"	18
	• 026-08	• "talking about lots of different stuff in just a little paragraph"	18
29	• 017-08	• all of the responses can be right	11
	• 026-08	• confusion caused by the list of steps (1,2,3,4,5) AND M/C responses (A,B,C,D)	00
30	• 001-11	• "chart difficult to pull information from"	15
	• 006-11	• responses C and D could both be right AND extraneous/confusing information in response D	11
	• 017-08	• words "safe level" are unclear	24
	• 023-08	• the chart is confusing	15
	• 025-08	• responses C and D could both be right	11