University of Alberta

Evaluating the Consistency of Verbal Reports and the Use of Cognitive Models in Educational Measurement

by

Xian Wang

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

Spring 2011
Edmonton, Alberta

Dedication

I would like to dedicate this work to my parents, Jiamin Wang and Minjuan Zhou. Their unconditionally support and encouragement are the inner strength to keep me moving forward.

Abstract

In the field of psychology, verbal reports are commonly used as a data source to explain human information processing. To date, few studies have investigated the accuracy of verbal reports for providing information on students' reasoning and problem solving on educational tasks. The purpose of this study is to evaluate the consistency of verbal data, as well as the effects of student achievement, interviewer knowledge level, and item difficulty on the consistency of verbal reports. Seventy-one Grade 12 students from two high schools provided verbal responses to 15 multiple choice test items from the Alberta Pure Mathematics Diploma Examination. Results indicate higher-achieving students demonstrate greater consistencies in verbal reports than moderate achieving students. The implications of the results are discussed and the limitations of the present study are also presented.

Acknowledgements

I would like to thank a number of people who have helped me throughout my time here at the University of Alberta. First, I would like sincerely thank my supervisor Dr. Jacqueline Leighton for offering me her support and guidance. As my mentor, Dr. Leighton provided guidance for me throughout my masters program in addition to overseeing the completion of my thesis. She provided me with many opportunities to gain experience and training to become a researcher, gave invaluable insights in the joys of writing, and most importantly, she encouraged me and kept me on task when I needed it the most. Without her patience and support, I would not have been able to complete this document.

I would like to thank Dr. Mark Gierl and Dr. Rebecca Gokiert for agreeing to be on my committee and provide comments for my thesis. I would also like to thank the professors in the Centre for Research in Applied Measurement and Evaluation (CRAME), particularly Dr. Todd Rogers, and Dr. Ying Cui for imparting knowledge of educational testing. I would also like to thank my colleagues in CRAME, Jiawen Zhou, Oksana Babenko and Hollis Lai, for their encouragement and helpful discussions associated with my study.

Finally, I would like to thank my family and friends. I am grateful to my parents Jiamin Wang and Minjuan Zhou for their understanding and support throughout my time in Canada. I am also grateful to have my friends, Hanhan Xue and Lei Jiang, for their company throughout this enlightening endeavour.

Table of Contents

List of Tables

List of Figures

CHAPTER I - BACKGROUND

Introduction

Education is always a major issue for government. The challenge of preparing children for success requires considerable investment in education, even in a time of an economic recession. In its 2010 Education Budget, the Government of Alberta stated that investment in Alberta's Kindergarten to Grade 12 education system will reach $6.3 billion--an increase of 0.8% over the previous year. In the budget for Basic Educational Programs, including Provincial Achievement Tests (PATs) (Alberta Education, 2010) and Diploma Exams, the amount of funding increased 1.6% to a total of $99 million in 2010. In the USA, in relation to the No Child Left Behind (NCLB) Act of 2001, the United States Department of Education highlighted the claim that educational testing should be used to measure student success in learning (USDE, 2004). The federal budget for the 2009 fiscal year saw an increase of 41% in spending for education since 2001, reaching a total of $24.5 billion (USDE, 2009). Of those funds, $409 million were spent on the State Assessment Grant (USDE, 2009). These government investments suggest that not only is education a worthy of investment, but educational testing is specifically considered a worthy investment as well.

Educational testing is an integral part of the educational process. Popham (2000) suggested that educational testing be described as the process of using students' responses to stimuli (i.e., educational tasks or items) to generate information or inferences about what students know, can do, and even how they feel. As students learn within an educational system, most notably a classroom,

testing in the educational context is often used as the way to determine whether students have obtained content mastery associated with a set of expected knowledge and skills. As a result, consequences from educational testing have an impact on student learning. As stated in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999), well-constructed and valid tests need to be used as a basis for making important, high-stakes decisions about students, such as classifying them as requiring extra-help and determining other options for their future education. Therefore, educational testing needs to be administered carefully, as no single test is adequate to measure everything that students know and can do.

The United States Department of Education claimed in their report on the NCLB of 2001 Act that "… tests will give teachers and principals information about how each child is performing and [tests will] help them to diagnose and meet the needs of each student"(USDE, 2004, 3rd paragraph). If tests are to accomplish this purpose—of providing information about the knowledge and skills students have acquired in the classroom and in their educational experiences generally—then tests must be designed and developed in accordance with information about what is known about how students acquire information and apply that information. Given that testing is expected to provide important information to guide serious decisions about student learning, it is crucial that research is completed to ensure that testing does provide decision-makers with the most accurate and reliable information about students' knowledge and skills.

Advances in research conducted in educational measurement have led researchers to become increasingly interested in improving the validity of inferences made from educational tests (Embretson, 1999; Ferrara et al., 2004; Leighton, Gierl, & Hunka, 2004; Mislevy, 1996). A growing number of researchers and practitioners are calling for the combined use of cognitive psychology and educational measurement to enrich investigations of testing, the development of test items, and the inferences made about students from their test results (Embretson, 1998; Gierl, Cui, & Hunka, 2008; Leighton et al., 2004; Mislevy, 2006; NRC, 2001; Pellegrino, Baxter, & Glaser, 1999; Snow & Lohman, 1989). In doing so, these researchers have indicated a need for making inferences about students' knowledge and skills, and generating information about students' internal mental processes to produce clear recommendations about the content knowledge and usable skills that are available to students. Understanding the mental processes students select, and apply in problem-solving is the essential feature of modern validity theory (Kane, 2006; Leighton & Gierl, 2007b; Leighton & Gierl, in press).

Recently, test developers, who are the "person(s) or agency responsible for the construction of a test and for the documentation regarding its technical quality for an intended purpose" (AERA, APA, NCME, 1999, p. 183), are meeting the challenge to develop tests that provide information beyond what traditional tests have done. Compared to traditional tests, which simply generate a rank order of where students stand or simply conclude the students have performed well or not (NRC, 2001), new types of educational tests are expected to provide students,

teachers, parents, and other stakeholders with more specific information about: (a) why students perform poorly and (b) how tests can be optimized to improve students' learning. That is, the new objective for test developers is to generate educational tests that "move us forward" in our understanding of students' cognitive skills, namely, their problem-solving strengths and weakness (Leighton & Gierl, 2007a, 2007b). These new tests provide diagnostic information about students in order to identify "both appropriate content and the types of learning activities that will help a student attain the learning target" (Nitko & Brookhart, 2007, p. 11). *Cognitive models of learning* can be used to define such diagnostic information by illustrating the knowledge and skills that underlie successful or competent performance in the content domain of interest (Leighton, 2004; Leighton & Gierl, in press; NRC, 2001). Research has suggested that tests developed from such cognitive models of learning have the potential to fulfill this objective in providing diagnostic information about examinees' cognitive strengths and weaknesses (Leighton, 2004; Leighton & Gierl, 2007a; NRC, 2001; Snow & Lohman, 1989).

In order to develop new, diagnostic tests, the cognitive models used must be supported with empirical evidence of human information processing (Leighton, 2004). Verbal reports are considered to be appropriate and are widely used to collect information on human information processing in the field of psychology (Ericsson & Simon, 1993) and educational measurement (Leighton & Gierl, 2007b). As a type of data that provides empirical evidence on human information processing, verbal reports are usually collected using think-aloud interviews

(Ericsson & Simon, 1993). During a think-aloud interview, the investigator asks the student to articulate his or her thoughts when solving a particular task and after the student has solved the task (Ericsson & Simon, 1993; Leighton & Gierl, 2007b). After verbal reports are collected, protocol analysis or verbal analysis can be used to analyze the verbal report data for the purpose of generating and/or validating cognitive models of learning (Leighton & Gierl, 2007b). When using verbal reports in educational testing research, special consideration must be given to the procedures used to collect verbal reports, as these procedures may influence the accuracy and consistency of the reports, and, in turn, the soundness of the inferences made about students' knowledge, skills, and problem solving (Ericsson & Simon, 1993). The present study is designed to evaluate one aspect of the procedures used in the collection of verbal reports, namely the consistency between concurrent and retrospective reports collected under different interviewer conditions, in order to improve the procedures used to collect verbal data and the development of cognitive models of learning for educational tests.

The paper contains four sections. The first section contains an overview of the literature, including an introduction to cognitive models and verbal reports. The second section provides details about the present study, including data collection, research design, and data coding procedures. The third section presents the results of the study. The fourth section concludes with a summary, discussion of results and limitations of the present study.

Literature Review

Cognitive Models and Educational Measurement

Students' academic achievement cannot be evaluated directly by means of observation, but can be measured indirectly from their responses to educational test items. In other words, by administering tests to students, educators can make inferences about students' knowledge and skills from their responses, in order to reach the goal of evaluating students' learning and achievement (Leighton & Gierl, 2007a; NRC, 2001). Traditionally, classroom tests have been used to evaluate students' learning and achievement. In addition to classroom tests, large-scale tests such as the SAT (Scholastic Aptitude Test) (College Board, 2010), the SAIP (School Achievement Indicators Program; now known as the PCAP or Pan-Canadian Assessment Program) (CMEC, 2007), the GRE (Graduate Record Examination) (ETS, 2010), and PISA (Programme for International Student Assessment) (OECD, 2007), are also administered to fulfill similar purposes.

When educators draw inferences about a student's performance on an educational task, whether on classroom tests or large-sale tests, two assumptions are implicitly made. The first assumption is that test items are constructed to measure specific skills and knowledge mastery. The second assumption is that students responding correctly to such questions have mastered the expected knowledge and skills, and have thought about the problem correctly to produce the keyed answer (Leighton & Gierl, 2007a). Recently, however, many studies have revealed that these assumptions may be not tenable in some cases (e.g., Gierl,

1997; Leighton & Gierl, 2007a; Leighton & Gokiert, 2005; Poggio, Clayton, Glasnapp, Poggio, Haack, & Thomas, 2005; Rogers & Yang, 1996). For example, Rogers and Yang (1996) found that test-wise strategies can enable students to generate correct responses without the prerequisite content knowledge. Moreover, when students respond incorrectly to test items, limited inferences can be made about the students, except to say that a student has performed poorly in a particular content domain (Leighton & Gierl, 2007b; NRC, 2001).

Increasingly, educational measurement specialists are unsatisfied with the limited and general information provided by classroom and large-scale tests (e.g., AERA, APA, NCME, 1999; Hamilton, Nussbaum, & Snow, 1997; Koretz & Hamilton, 2006; NRC, 2001). One of the arguments that has been made is that general information about student performance in reference to other students has limited value in improving teaching and learning (NRC, 2001). By providing more specific information as to how students have performed, educational tests, including both classroom and large-scale tests, should be better able to direct teaching for improving student learning.

Most classroom and large-scale tests are currently developed from test specifications or blueprints (Leighton & Gierl, 2007a). Although test specifications illustrate the knowledge and skills desired for measurement within the content domain, they describe these skills and knowledge using a large grain size, which restricts the measurement of students' achievement to a general level and forces inferences to be kept at a general level as well (Leighton & Gierl, 2007a). Therefore, tests developed from test specifications cannot provide

specific diagnostic information of examinees' cognitive strengths and weakness (Leighton & Gierl, 2007a). Moreover, there is little empirical evidence to show that the thinking processes examinees have *actually* used to answer the test items are aligned to the content and skills illustrated in the test specifications (Leighton & Gierl, 2007a, 2007b; Nichols, 1994). Hence, although test specifications could be seen as representing a model of cognition, most, if not all of the knowledge and skills included in test specifications, represent a *hypothesized* model of cognition (Leighton & Gierl, 2007a) in need of empirical support. Tests developed from test specifications can be used as measurement instruments of expected knowledge and skills in students, but whether or not students do, in fact, employ the knowledge and skills outlined in test specifications needs to be verified.

In order to determine the specific thinking process examinees use to solve tasks and generate answers, so as to identify their cognitive strengths and weaknesses, empirically-based cognitive models of learning can be used to develop tests (Leighton & Gierl, 2007a, 2007b; Nichols et al., 1995; Norris et al., 2004; NRC, 2001). The use of cognitive models to represent human information processing was first implemented in the field of computer science, and then applied in the field of cognitive psychology (Leighton & Gierl, 2007a, 2007b). Leighton and Gierl (2007a, p. 6) defined the term cognitive model in the context of educational measurement as a "simplified description of human problem solving on standardized educational tasks, which helps to characterize the

knowledge and skills examinees at different levels of learning have acquired and to facilitate the explanation and prediction of students' performance."

As mentioned previously, most classroom and large-scale tests are currently developed from test specifications, which outline the knowledge and skills a student is expected to use to solve test items (Gierl, 1997; Leighton & Gierl 2007a). However, according to Leighton (2004), a cognitive model of test specifications is only one type of model among three possible types of cognitive models that could be used to design a test. A second type is a cognitive model of domain mastery, which is usually used to design curriculum-based tests (Leighton, 2004; Leighton & Gierl, 2007a). Tests constructed from a cognitive model of domain mastery cannot support inferences about examinees' thinking processes because the model that underwrites the test focuses on examinees' mastery of knowledge and skills at a behavioral level (Leighton & Gierl, 2007a). Similarly, tests developed from a cognitive model of domain mastery lack empirical support because these cognitive models are not developed based on studies of student thinking. Thus, tests designed from cognitive models of domain mastery are relatively weak at providing diagnostic inferences about examinees' cognitive strengths and weakness (Leighton & Gierl, 2007a).

The third type of model identified by Leighton (2004) is a cognitive model of task performance. In educational measurement, such a model reflects a fine-grain description of the knowledge and skills examinees have been found to use to solve educational tasks in a specific content domain (Leighton, 2004; Leighton & Gierl 2007a, 2007b). According to Leighton and Gierl (2007b), cognitive models

of task performance illustrate the declarative knowledge and procedural skills used to transform and manipulate test item information to generate a response (Lohman, 2000). By illustrating the detailed thinking processes underlying the knowledge and skills a student is expected to use to solve an educational task, a cognitive model of task performance explicitly maps examinees' problem solving procedures in a step by step manner (Leighton & Gierl 2007a). Therefore, tests designed from a cognitive model of task performance are expected to provide more specific information about examinees' thinking and learning processes. Tests developed from cognitive models of learning are known as *cognitive diagnostic assessments* (CDA) (Leighton & Gierl, 2007b).

Unlike cognitive models of test specifications and domain mastery, which are usually experimentally unsubstantiated, cognitive models of task performance are usually supported by empirical evidence of examinees' thinking processes (Frederiksen et al., 1990; Leighton, 2004; Leighton & Gierl, 2007b; Nichols, 1994). Therefore, only substantiated cognitive models of task performance can inform strong inferences about what students know and can do in response to educational items or tasks (Leighton, 2004). Consequently, the development of CDAs is often based on cognitive models of task performance. Further, validating test inferences based on cognitive models of task performance can be completed using a variety of approaches (Leighton, 2004); one approach of interest for the present study is the use of *verbal reports*.

Verbal Reports

There are a variety of procedures that could be used to investigate human information processing (e.g., response latencies, eye fixations, and extended essays; see Leighton & Gierl, 2007b, Lohman, 2000). However, verbal reports are considered an appropriate data source to reveal information about how examinees think through tasks (Hamilton, Nussbaum & Snow, 1997; Leighton, 2004; Leighton & Gierl, 2007b; Norris, 1994). Specifically, verbal reports can provide information about students' reasoning and problem solving on educational tasks (e.g., Hamilton et al., 1997; Leighton & Gierl, 2007b).

A verbal report is an individual's description of his or her internal processes for solving a problem (Ericsson & Simon, 1993). Verbal reports are often used to develop cognitive models (Ericsson & Simon, 1993; Ericsson, 2006). For example, Leighton (2004) indicates that collecting verbal reports from students on a given task of interest can be used to develop a cognitive model of task performance. Protocol analysis (Ericsson & Simon, 1993) or verbal analysis (Chi, 1997) can then be used to analyze the verbal reports. Protocol analysis is often used to identify a linear sequence of thought processes, while verbal analysis is often used to identify the knowledge representations a student has developed about a task (Chi, 1997; Leighton & Gierl, 2007b).

Verbal reports are collected using standardized procedures, and are usually conducted using one-to-one interviews. When verbal reports are collected, examinees are asked to think aloud as they solve a task. That is, examinees are asked to verbalize every thought that comes to mind as the problem is being

solved (Ericsson & Simon, 1993; Kuusela & Paul, 2000; Leighton, 2004; Leighton & Gierl, 2007b; Taylor & Dionne 2000; Van Gog et al., 2005b). This type of reporting is called a *concurrent verbal report*. Another type of verbal report that is collected immediately following the concurrent report is called a *retrospective verbal report*. During the retrospective report, examinees are instructed to recall their thought process after they have solved the task (Ericsson & Simon, 1993; Leighton & Gierl, 2007b; Taylor & Dionne, 2000; Van Gog et al., 2005b). In the paragraphs that follow, concurrent and retrospective verbal reports are described in greater detail.

Concurrent and Retrospective Verbal Reports

In concurrent reporting, students are instructed to think aloud as the activity occurs, creating a direct record of their thoughts during the time the task is being solved (Ericsson & Simon, 1993; Leighton & Gierl, 2007b; Van Gog et al., 2005b). That is, concurrent reports provide *on-line* information about a student's processed information in the form of the response they provide to a task, and also how such information was manipulated prior to arriving at the response (Ericsson & Simon, 1993; Kuusela & Paul, 2000; Leighton, 2004; Leighton & Gierl, 2007b). From a cognitive psychology perspective, concurrent reports reflect the information stored in working memory (WM) during problem solving (Ericsson & Simon, 1993; Taylor& Dionne, 2000; Van Gog et al., 2005b).

In retrospective reporting, students are asked to think back and recall their thoughts about how they solved the task. Retrospective reports are often collected

immediately after students have completed the task and produced a final answer (Ericsson & Simon, 1993; Leighton & Gierl, 2007b; Taylor& Dionne, 2000; Van Gog et al., 2005b). That is, retrospective reports contain *off-line* information on what students remember in solving the task. In the context of cognitive psychology, retrospective reports measure information retrieved from short-term memory (STM) and long-term memory (LTM) depending on the length of time between when the task was solved and the retrospective report was initiated (Camps, 2003; Ericsson & Simon, 1993; Taylor& Dionne, 2000; Van Gog et al., 2005b). The longer the lag between the completion of the task and the collection of the retrospective report, the more likely it is that the retrospective report reflects contents of LTM and not STM.

Concurrent reports are expected to reveal the sequence of students' problem solving in WM (Leighton & Gierl, 2007b). However, because concurrent reports are expected to measure the contents of working memory, there are two reasons why concurrent reports may not be complete or accurate. First, only cognitive processing that is consciously noted can be verbalized. Therefore, subconscious information derived from automatic processing, which does not register in WM, may not be verbalized and reported (Leighton, 2004; Taylor & Dionne, 2000). Second, human WM is known for its size limitation (Miller, 1956). As a result, the limited processing capacity of WM may restrict students from reporting the full extent of the thinking processes used to solve an educational task (Kuusela & Paul, 2000; Taylor& Dionne, 2000). Related to the constraints associated with the memory location tapped by concurrent reports, researchers

have expressed concerned with the accuracy and validity of concurrent verbal reports. For example, Wilson (1994) pointed out that providing a verbal report while solving a task may alter the course of thinking or the application of cognitive resources, which in turn could change the underlying processes used in problem solving (Kuusela & Paul 2000; Leighton & Gierl, 2007b). This is known as *reactivity* (Ericsson & Simon, 1993; Leighton & Gierl, 2007b) and could influence the accuracy, and in turn, validity of the verbal report. Another common concern expressed by researchers is that verbal reports elicited from students may not indicate the thinking processes they *actually* used to solve the task (Ericsson & Simon, 1993, Leighton & Gierl, 2007b; Payne et al., 1978). This is referred to as *non-veridicality* (Ericsson & Simon, 1993; Leighton & Gierl, 2007b) and it could also influence the accuracy, and in turn, the validity of the verbal report.

Consistency is also an issue of concern when collecting verbal reports. Retrospective verbal reports serve as a useful method to monitor (and verify) the accuracy of the problem solving reported in concurrent reports (Leighton & Gierl, 2007b). Retrospective verbal reports reflect the contents of STM and LTM depending on the length of time between the end of the concurrent report (solution of the task) and the beginning of the retrospective report (Ericsson & Simon, 1993; Leighton & Gierl, 2007b; Taylor& Dionne, 2000; Van Gog et al., 2005b). For example, Ericsson and Simon (1993) claimed that a person can remember his or her thinking processes accurately and congruently only if the person completed the task within a short duration (0.5 – 10 seconds) before commencing the retrospective report. According to Ericsson and Simon (1993),

retrospective reports that are collected long after the task has been solved will tend to deviate from the contents of concurrent verbal reports in terms of accuracy and completeness. In other words, depending on the procedures followed during the collecting of verbal reports, retrospective reports may not be consistent with concurrent reports and, therefore, not be useful in informing the contents of students' thinking processes.

Previous Studies on Consistency of Verbal Reports

A major concern for researchers who employ verbal reports is the accuracy and consistency of the thinking processes students report using to solve a problem (Leighton, 2004). Although both concurrent and retrospective reports are collected to provide a "double measure" on students' thinking processes, the consistency between concurrent and retrospective reports has not been the subject of study in educational measurement. Several researchers in fields other than educational measurement have conducted empirical studies on the consistency, or similarities and differences, between concurrent and retrospective verbal reports and the results have been mixed. In a study conducted by Taylor and Dionne (2000), they found that the codes assigned to concurrent reports of the actions taken during problem solving (e.g., "What I did here was sketch out who the groups are and what their interests are…") were significantly higher in number than in retrospective reports. Likewise, the codes assigned to metacognitive processes in retrospective reports were significantly higher in number than in concurrent reports. Taylor and Dionne concluded that concurrent report data primarily provide information on the cognitive actions taken during problem

solving, whereas retrospective report data provide information primarily on students' use of conditional knowledge about cognitive actions, and beliefs, or metacognition of the conditions that warrant particular cognitive actions (or problem solving processes) to be executed. Taylor and Dionne (2000) concluded that the benefit of comparing concurrent and retrospective reports for their consistency rests with enhancing the validity of verbal data generally because retrospective reports often make explicit what is left implicit in a concurrent report.

Similarly, Kuusela and Paul (2000) compared the effectiveness of concurrent and retrospective data for revealing human decision-making processes. These investigators found that concurrent and retrospective reports were not always consistent. For example, the total number of task-relevant statements was higher in concurrent reports than in retrospective reports. However, the average number of protocol segments (larger units than task statements) was significantly higher in retrospective reports than in concurrent reports. Kuusela and Paul (2000) did find, however, that participants showed a consistent understanding of the instructions used during the concurrent and retrospective interviews. They concluded that concurrent reports yielded more detailed information about participants' decision making processes, while retrospective reports were found to be more suitable in revealing information about participants' decision outcomes—that is, participants' final answer choices. These results suggested that the information contained in concurrent and retrospective verbal reports may not be necessarily identical but they may be complementary. Therefore, the methods

used to conduct verbal reports should be carefully chosen depending on the type of information the investigator wishes to obtain.

In another study, Van den Haak, De Jong and Schellens (2003) presented an experimental study on an online library catalogue to compare concurrent and retrospective reports for usability testing. They noted that on a global level, concurrent and retrospective reports were not significantly different in the number and types of problems detected. However, the two methods differed significantly with regards to the manner of problem-detection; that is, concurrent verbal reports revealed more problems by means of observation alone, while retrospective verbal reports resulted in significantly more problems detected by means of explanation. The study revealed no significant differences between the reports in terms of time for task completion. However, the researchers suggested the study had a drawback in that the tasks used in the experiment were too difficult for the participants. When tasks are too difficult or too easy, the contents of verbal reports may not represent the thinking processes used to solve the tasks.

According to Ericsson and Simon (1993), when tasks are too difficult, participants may stop verbalizing because the cognitive load in WM is too heavy. Previous studies have found that novel tasks of moderate difficulty are the most appropriate tasks to be used for eliciting verbal reports (Afflerbach & Johnson, 1984; Ericsson & Simon, 1993; Leighton, 2004; Taylor & Dionne 2000). Likewise, when tasks that are too easy or too familiar for students are used to elicit verbal reports, students may solve these tasks using “automatic processing” or well-learned routines that essentially bypass the conscious control needed for

verbal reporting (Leighton, 2004). In other words, problem solving of easy tasks may occur so quickly so as to bypass conscious awareness in WM. As a result, students may not be aware of how they solved the task or may have difficulty in describing their problem-solving processes (Leighton, 2004; Wilson, 1994). Therefore, such easy tasks will yield poor verbal report data due to a lack of verbalization (Ericsson & Simon, 1993; Leighton 2004).

The possible reactivity and non-veridicality of verbal reports are serious issues in educational testing. If reactivity and non-veridicality exists during a think-aloud interview, the validity of the verbal report will be undermined as the reported contents may not be aligned with students' thinking processes (Leighton, 2009). Although Ericsson and Simon (1993) have developed standard procedures to collect verbal reports, these procedures may not always be followed by researchers. Ericsson and Simon (1993) put forward these procedures to minimize reactivity as well as non-veridicality. For example, when collecting a concurrent report, interruptions from the interviewer should be minimized, and standard probes and prompts should be used to remind the student to continue vocalizing his or her thoughts (Leighton & Gierl, 2007b). Further, when collecting retrospective reports, verbal data should be collected as soon as possible following the completion of the task to optimize retrieval from STM (Ericsson & Simon 1993; Presley & Afflerbach, 1995; Taylor & Dionne; 2000, Van Gog et al., 2005b). Also, in order to reduce anxiety and ensure the participants' understanding of the verbal report instructions, standard instructions should be followed, including a warm-up exercise that is usually conducted to familiarize

students with thinking aloud (Ericsson & Simon, 1993; Taylor & Dionne 2000; Van Someren et al., 1994). The full procedure for conducting verbal reports is outlined in Ericsson and Simon (1993).

Another issue that is often overlooked is the domain in which Ericsson and Simon (1993) conducted their review of verbal report data. Ericsson and Simon (1993) conducted a review of past studies and concluded that verbal reports collected using standard procedures did not alter the structure of thinking processes. However, most of the studies included in their review were conducted in psychological laboratories with psychological tasks and not educational test items (Leighton, 2010). Although Ericsson and Simon (1993) outlined specific procedures for the collection of verbal reports to ensure the accuracy and consistency of verbal reports, it has to be noted that Ericsson and Simon's procedures were designed for collecting verbal reports in response to psychological tasks, and not achievement test items or other potentially high stakes educational tasks (Leighton, 2010).

One of the studies included in Ericsson and Simon's review provides an exception. For example, Norris (1990) conducted a study on the effects of eliciting verbal reports in response to critical thinking test items. Norris used a complete randomized factorial design to investigate four slightly different ways of eliciting verbal reports from examinees as they worked on a particular multiple-choice critical thinking test. The four conditions included *think aloud*, *immediate recall*, *criteria probe*, and *principle probe*. Each condition reflected a different way of eliciting students' thoughts. For example, in the *think aloud* condition,

students were instructed to report all their thoughts while they were solving an item and then mark their answer on the answer sheet. In contrast, in the *immediate recall* condition, students were instructed to first mark the answer to an item on the answer sheet and, after they had marked their answer, verbalize why they had chosen the answer. The students who took a standardized paper-and-pencil test and did not verbalize their thoughts constituted the control group. Norris found that the collection of verbal reports using different methods did not alter students' thinking or performance on task. Norris concluded that elicitation of verbal reports on students' thinking did not influence the course of their thinking relative to the control group in the study. However, in the same study, Norris (1990) did find interviewer related effects. His results showed that when interviewer A conducted the interviews, male and female subjects performed equally well on a pencil-and paper test and on verbal reports. When interviewer B conducted the interviews, male subjects performed significantly better than female subjects. This result suggested that students' responses changed when facing different interviewers during the think-aloud interview. These findings suggest that students' thinking processing may change as a function of certain types of tasks (i.e., critical thinking test items) and different interviewers. Therefore, more research is needed to investigate the effects of test items and interviewer effects in think-aloud studies (Leighton, 2010).

To date, the accuracy and consistency of verbal data in achievement testing studies has not been investigated. When participants respond to psychological tasks, they are often told that there is no right or wrong answer.

However, participants know that there are right or wrong answers to achievement tests items. Many students become nervous or anxious when taking educational tests (Sawyer & Hollis-Sawyer, 2005). This nervousness or anxiety could interfere with students' thinking processes and in turn, the accuracy and consistency of verbal reports. Further, students' anxiety in the face of providing verbal reports to educational test items may be exacerbated if they believe they are being interviewed by an expert interviewer. An expert interviewer is in a position to evaluate students' thinking and responses, and judge whether the student is a "smart" person.

The Need for the Present Study

Verbal reports, both concurrent and retrospective, are often collected to provide evidence of students' thinking processes so as to develop and/or confirm cognitive models of learning for the development of CDAs. However, one of the important implications from the literature review is that the information collected from concurrent and retrospective verbal reports on problem-solving may not always be accurate or consistent. Although researchers in other disciplines such as psychology have investigated the data obtained from concurrent and retrospective reports, these studies have not been conducted with educational test items. Since there have not been studies focused on the accuracy and consistency of students' verbal report data in educational measurement, the present study will focus on investigating concurrent and retrospective reports from such a perspective. In particular, the present study will focus on the consistency of concurrent and retrospective reports when interviewer characteristics are manipulated, as well as

test item features. In addition, student achievement level was controlled because it was expected that students of higher ability might be more consistent in their verbal reports than students of lower ability.

It is important to note that the present study is part of a larger study that was funded by a standard research grant awarded to the supervisor of this thesis (Dr. Jacqueline Leighton) by the Social Sciences and Humanities Research Council of Canada (grant number 410-2007-1142). In this larger study many more dependent variables were examined such as the *accuracy* of test item answers in response to interviewer characteristics in the think-aloud interview, as well as the *familiarity* and *confidence level* of the students when solving test items. Measures of *meta-cognition*, *test anxiety*, and students' *perceptions of the think-aloud interview* were also administered to students to evaluate how they responded to interviewer characteristics in the think-aloud interview. Moreover, the *ability level* of the cognitive models students produced in their concurrent and retrospective verbal reports in response to interviewer characteristics and test item characteristics were evaluated. Readers interested in learning about the results of the larger study are referred to Leighton (2010, 2011).

CHAPTER II – METHODS

Purpose and Research Questions

The purpose of this study was to evaluate the consistency of verbal report data in providing information on students' cognitive models during problem solving. Specifically, this study investigated verbal data from student responses to 15 multiple-choice mathematic questions in order to explore the consistency of concurrent and retrospective reports. Moreover, this study also investigated the effects of student achievement, interviewer knowledge level, and item difficulty on the consistency of verbal reports. To summarize, this study addressed the following research questions:

1. Are concurrent and retrospective verbal reports consistent across math test items?
2. Is the consistency of concurrent and retrospective reports influenced by student achievement level? Specifically, between high achieving and moderate achieving students?
3. Is the consistency of concurrent and retrospective reports influenced by interviewer knowledge level?
4. Is the consistency of concurrent and retrospective reports influenced by item difficulty (i.e., easy, moderate, vs. difficult items)?

The following sections outline the methods used in the current study. First, the participants interviewed in this study are described by their demographic characteristics. Second, materials used in the interview are summarized. Third, the

experimental design of the study is explained. Fourth, the procedures used for the experiment are presented. Finally, the coding scheme of the interview data for statistical analysis is explained.

Participants

There were 71 participants involved in the study, which included 39 girls and 32 boys ($M = 17.14$, $SD = 0.61$) enrolled in a Grade 12 university-tracked pure mathematics course. All students come from two academic high schools in a medium-sized city and selected by their teachers as capable of verbalizing their thinking processes as they solved mathematics problems. Among these 71 students, 38% self-identified as Caucasian/white, 21% as Asian/Chinese, 16% as Filipino, 10% as South Asian or Southeast Asian, and the remaining 15% self-identified as Black, Korean, Latin American, Arab, or Other. Parental consent was sought for participation in the study and students were compensated with a gift certificate to purchase a book for the time they contributed.

Materials

Each of the 71 students participated in a think-aloud interview. Each student was presented with a booklet that contained five sections and a cover page. The students were asked to fill out their name, date, and the starting time of their interview on the cover page. The first section of the booklet included 15 mathematics items, presented separately on individual pages, in the domain of pure mathematics, accompanied with a familiarity scale and a confidence scale on

the other side of the page[1]. Familiarity and confidence scales were included so that students could rate how they felt about their responses. After completing each math item, students were given the following prompts: "using the scale below where 0% means 'not at all familiar' and 100% means 'absolutely familiar,' please circle a value indicating how familiar you are with the information in this item." After this, students also were prompted: "using the scale below where 0% means 'not at all confident' and 100% means 'absolutely confident,' please circle a value indicating how confident you are with the information in this item." The 15 test items were arranged by item difficulty: five easy items were presented first with one item on each of 5 pages, followed by five moderate items presented with one item on each of 5 pages, and five difficult items presented with one item on each of 5 pages. The second section contained a 20-item four-point Likert scale designed to evaluate students' metacogniton called the *Self-Assessment Questionnaire* (O'Neil & Abedi, 1996)[2]. The third section of the booklet contained a 20-item four-point Likert scale measure of test anxiety called the *Test Attitude Inventory* (Spielberger, etal., 1980). The fourth section contained a questionnaire made up of 8-items using a seven-point Likert scale to evaluate the perception of students toward think-aloud interviews. The last section of the booklet contained a 4-item questionnaire asking for students' demographic

[1] Students' responses to the familiarity and confidence scales are presented in Leighton (2010). The present study focuses on the consistency between the contents of concurrent and retrospective reports.

[2] Students' responses to the Self-Assessment Questionnaire, the Test Attitude Inventory, and an additional 8-item questionnaire are presented in Leighton (2010).

information. Students were requested to finish the sections one by one in the order presented and to not skip sections or look ahead to next sections.

The 15 multiple choice test items, used in the first section of the booklet, were selected from a set of released items belonging to the Alberta Pure Mathematics Diploma Examination. This Diploma Examination is a large-scale assessment of academic Mathematics achievement at the Grade 12 level in the province of Alberta. These items were chosen with the assistance of two experienced test developers who worked for the assessment branch of the provincial education ministry (Alberta Education) and who had experience in assessing Grade 12 students. As part of a large-scale assessment in pure mathematics, which was developed professionally in relation to the Grade 12 mathematics course curriculum, the content of these items was believed to adequately cover what the enrolled students should have learned in their Grade 12 pure mathematics course, including both trigonometry and calculus knowledge. All items had acceptable levels of item discrimination, and the items could generally be clustered into three levels of difficulty. The five easy items had p-values greater than 0.7, the five moderate items had p-values between 0.4 and 0.7, and last five difficult items had p-values less than 0.4. P-values indicate the proportion of examinees who answered the item correctly.

Design

The experimental design included between-subject and within-subject variables. Two interviewer variables were manipulated – gender and knowledge

level were between-subject variables[3]. One item variable was manipulated – item difficulty was a within-subject variable. A graphic of the design is shown in Figure 1. Figure 1 only shows interviewer knowledge level as this is the between-subject variable of interest in the present study. Two interviewers were hired to conduct all 71 interviews. The two interviewers included two graduate student research assistants who were of approximately the same age, one male and one female. They were both Caucasian, spoke English as their first language, and conducted interviews in casual dress and demeanour. Both of them were trained with the same think-aloud instructions outlined by Ericsson and Simon (1993) and conducted all interviews using verbatim text and instructions.

The between-subject variable, *interviewer knowledge level*, involved three levels: high knowledge in mathematics (i.e., expert condition), low knowledge in mathematics (i.e., novice condition), and no information about knowledge condition (i.e., control condition). Interviewer knowledge level was manipulated by varying a subset of instructions used to conduct the think aloud interviews. The following instructions were used to conduct the think aloud interviews. Section 2 of the instructions was manipulated depending on the condition of interviewer knowledge level being presented to students:

1. *Thank you for taking part in today's study involving math diploma test questions. I'll also be asking you to complete two short questionnaires—one involving strategies you used to solve the test questions and another*

[3] Although interviewer gender was manipulated, it is not examined in the present study because no interviewer gender effects' were found on students' response accuracy in a related study (Leighton, 2010).

on test attitudes. Please know that your help today is completely voluntary and you are free to go at any time. Now, before I explain what we will be doing today, let me introduce myself. The study will take no more than an hour.

2. *My name is _____ and I'm from the University of Alberta. And I'll be conducting the interview today* [control condition].
3. *So, now let me tell you about the study you're involved with today. I will read this because it is important and I want to make sure everyone gets the same instructions.*

   In this study we are interested in what goes through your mind or what you think about when you find answers to questions in math. In order to do this I'm going to ask you to THINK ALOUD as you work on the problems given. What I mean by think aloud is that I want you to tell me EVERYTHING you are thinking from the time you first see the question until you give an answer. I would like you to talk aloud CONSTANTLY from the time I present each problem until you have given your final answer to the question. I don't want you to try to plan out what you say or try to explain to me what you are saying. Just act as if you are alone in the room speaking to yourself. It is most important that you keep talking. If you are silent for any long period of time I will remind you to talk. Do you understand what I want you to do? After you finish talking aloud as you solve the problem I will ask you to tell me how you REMEMBERED solving the problem. I know this might seem like I'm repeating myself but I

want to make sure I understand how you solved it. I will audio record our session because I want to get an accurate record of your think aloud reports. Please know that all the information you share with me today will be kept private. Do you have any questions?

Next, students received the following instructions:

4. *Now we will begin with a practice problem.*

   "What is the result of multiplying 24 x 36?"

   Next, students were given time to solve the problem and think aloud.

   Now I want to see how much you can remember about what you were thinking from the time you read the question until you gave the answer. I am interested in what you actually can REMEMBER rather than what you think you must have thought. If possible I would like you to tell me about your memories as they occurred while working on the question. Please tell me if you are uncertain about any of your memories. I don't want you to work on solving the problem again, just report all that you can remember thinking about when answering the question. Now tell me what you remember.

During the interview, probing statements were used for eliciting students' concurrent verbal reports. An example of a probing statement included: "*Remember I'm just interested in your thoughts, please continue talking."* For eliciting retrospective verbal reports, example of such probes included: *"Please tell me all that you can remember about how you solved this problem," "Let's go on to the next question (if student was finished with item)."*

Interviewers were trained to articulate their knowledge level naturally at section number 2 in these instructions for students. As shown in the instructions presented previously, the control condition was manipulated with the following text:

> *My name is _____ and I'm from the University of Alberta. And I'll be conducting the interview today.*

The expert interviewer knowledge condition, at section number 2, was presented with the following text:

> *My name is _____ and I'm from the University of Alberta. My area of expertise is in Mathematics and I've been interested in how students solve problems for many years. And I'll be conducting the interview today.*

Likewise, the novice interviewer knowledge condition, at section number 2, was presented with the following text:

> *My name is _____ and I'm from the University of Alberta. My area of expertise is not in Mathematics but I've been interested in how students solve problems for many years. And I'll be conducting the interview today.*

This was the only part of the instructions that differed across the experimental conditions of interviewer knowledge level. All the other instructions were identical across the three knowledge level conditions.

The 71 students were randomly assigned to one of the six experimental conditions (see footnote 4 and Figure 1), ensuring student gender was

approximately equally distributed across the six conditions. All 71 students responded to all 15 math items and thus all levels of item difficulty.

Procedure

All 71 participants were randomly assigned to one of six experimental conditions as a result of crossing interviewer gender (male vs. female) with interviewer knowledge level (expert, novice or control; see footnote 4). A random sampling schedule was used to make sure that each interviewer was conducting an approximately equal number of interviews with both boys and girls. The two interviewers conducted all 71 think-aloud interviews. The think-aloud interviews were conducted in a quiet room at the students' school, and each student participated in the interview individually. At the start of the think-aloud interview, an interviewer presented a booklet to a student, introduced himself or herself, as well as provided an explanation of the study's objective. The students were asked to provide not only concurrent but also retrospective verbal report data as they completed the 15 mathematics test items. As mentioned previously, interviewers reminded the participants during the interview to think-aloud. Concurrent interview probes included "*Remember I'm just interested in your thoughts, please continue talking,"* and retrospective probes included *"Please tell me all that you can remember about how you solved this problem," "Let's go on to the next question (if student was finished with item)."* The interviews were completed approximately within 45 minutes to one hour, and the entire think-aloud interview was audio recorded by using a digital sound recording device. After the interview was completed, students completed the questionnaires (see footnote 3). After

students completed the questionnaires, they were thanked and presented with the book certificate as compensation for their time and effort.

Coding of Think-aloud Interviews for Cognitive Modeling

Concurrent and retrospective verbal reports were coded according to five categories of cognitive models. Two test developers (KM and HR) were hired from Alberta Education to independently indentify the knowledge and skills associated with each math item, and then develop cognitive models for each of 15 math items. Both developers possessed extensive experience in developing large-scale items in pure mathematics. Given the consideration that low ability students may not correctly solve the items as well as articulate their thinking clearly, the test developers were hired to identify the knowledge and skills that moderate ability and high ability students were likely to use to correctly respond to each of 15 test items. Developers could indicate the same cognitive model for both moderate and high ability students if the developers considered the same knowledge and skills would be used by both types of students. After identifying the knowledge and skills likely to be used to solve each of the 15 items, the test developers were asked to assemble the knowledge and skill components into a cognitive model of learning in pure mathematics. The cognitive models were diagrammatically represented as flowcharts (Please see Appendix A and Appendix B for all cognitive models per item). Thus, each item was associated with four cognitive models – a moderate-ability model developed by KM, a high–ability model developed by KM, a moderate–ability model developed by HR, and a high–ability model developed by HR. Given that there were 4 cognitive models

developed for each item, and there were 15 items, 60 models were developed all together.

The cognitive models of learning developed for each item were used to classify students' verbal reports. Two graduate students, who were blind to the experimental study conditions, were trained as raters to classify the verbal reports. Working independently, the concurrent and retrospective verbal reports for each test item were classified by the raters into one of the four cognitive models of mathematical thinking developed for a particular item (i.e., test developer KM's moderate-ability cognitive model, KM's high-ability cognitive model, test developer HR's moderate-ability cognitive model, or HR's high-ability cognitive model). If the students' verbal reports did not reflect any of the cognitive models developed for a particular item (i.e., the student indicated no idea as to how to solve the problem or was just guessing the answer or used another model that did not match the test developer's models), then the response was classified as no model. At the start, the two raters were trained using three student interviews. Cohen's kappa was calculated as an index of inter-rater agreement to evaluate the agreement in classification. The higher the index, the more agreement achieved by raters. A kappa value of .81 was obtained for the inter-rater agreement on the three interviews. After discussing and resolving disagreements, a kappa value of 1.0 was obtained in the classification of verbal reports into one of five categories of models. In order to minimize any misclassification, however, both raters listened to all 71 interviews and conducted all 2130 (71 participant x 15 items x 2 type of reports = 2,130) item to model classifications. The categorizing of all

concurrent and retrospective verbal reports according to one of the 5 categories was quantified as follows: Verbal reports classified into test developer KM's moderate-ability cognitive model were assigned a value of 1; verbal reports classified into test developer KM's high-ability cognitive model were assigned a value of 2; verbal reports classified into test developer HR's moderate-ability cognitive model were assigned a value of 3; and verbal reports classified into test developer HR's high-ability cognitive model were assigned a value of 4. Lastly, if verbal reports reflected no cognitive model, the reports were assigned a value of 5. It is important to note that a student could have solved the item correctly but have his or her verbal report not match any of the cognitive models of moderate or high ability developed by the two test developers. In this case, the students' verbal report was designated as showing no cognitive model, namely, it was assigned a value of 5. To summarize, the coding is shown in Table 1.

Coding of Think-aloud Interviews

The objective of the present study was to evaluate the consistency of concurrent and retrospective verbal report data in providing information on students' cognitive models during problem solving. Specifically, this study investigated the effects of student achievement, interviewer knowledge level and item difficulty on the consistency of verbal reports for providing information on students' cognitive models during problem solving.

Students' concurrent and retrospective verbal reports were categorized into one of five model categories. After students' verbal reports were categorized, four

distinct coding procedures were followed, from the most restrictive to the least restrictive. To summarize and analyze results, the coding is shown in Table 2. The first two columns of Table 2 illustrate the permissible categories for coding students' verbal reports. For example, the first row of the first and second column in Table 2 illustrates that a student's concurrent verbal report could be assigned a value of "1" (indicating that it has been categorized according to KM's moderate-ability model) and his or her retrospective report could be assigned a value of "1"again (KM's moderate-ability model). The second row indicates that a "1" could be assigned to a student's concurrent report and a "2" could be assigned to the student's retrospective report. The remainder of the rows under columns 1 and 2 can be interpreted in like fashion. The third column of Table 2 illustrates the first coding scheme of the categorization of reports. The first coding scheme reflected the strictest coding for consistency between concurrent and retrospective reports. If, for a given item and a given student, the concurrent report and retrospective report showed identical cognitive model classification, then a value of 1 was assigned for consistency, otherwise a zero. This type of coding was used to determine whether students' concurrent and retrospective reports were completely identical in the cognitive model reflected in their thinking processes.

The second coding scheme reflected a less strict coding for consistency in concurrent and retrospective reports. If, for a given item and a given student, the concurrent report and retrospective report were categorized into a *similar ability-level of model*, then a value of 1 was assigned for consistency, otherwise a zero. This type of coding was used to determine whether students' concurrent and

retrospective reports matched in terms of ability—regardless of whether the reports were matched to distinct cognitive models designed by the different test developers. For example, if a student's concurrent report was matched to KM's high ability model and his or her retrospective report was matched to HR's high ability model, a value of 1 was assigned for consistency, otherwise a zero. The third coding scheme reflected a similar level of rigor as the second procedure. If for a given item and a given student, the concurrent and retrospective report were categorized into one of the two models designed by a single test developer, then a value of 1 was assigned, otherwise a zero. This type of coding was used to determine whether students' concurrent and retrospective reports matched in terms of the knowledge and skills each of the test developers considered relevant for success on the item. The last procedure was the least restrictive of all. If for a given item and a given student, *both* concurrent and retrospective reports showed a categorization into any kind of cognitive model (i.e., a 1, 2, 3, or 4 classification), then a value of 2 was assigned (i.e., cognitive models present for both concurrent and retrospective verbal reports). If the concurrent *or* retrospective showed a categorization into any kind of cognitive model, then a 1 was assigned (i.e., cognitive model present in only one of the verbal reports). If neither the concurrent nor the retrospective report showed a categorization of cognitive models, then a zero was assigned. This type of coding was used to evaluate the degree to which students' concurrent and retrospective reports reflected full use of cognitive models or even partial use of models.

Summing Consistency Values

Consistency values were summed for all 15 items for each coding scheme. For the first, the second, and the third coding schemes, the highest consistency value that could be obtained for any student was 15 if a student earned a "1" for each item. The lowest consistency value that could be obtained for any student was 0 if a student earned a "0" for each item. For the last coding procedure, the highest consistency value that could be obtained for any student was 30 if a student earned a "2" for each item. The lowest consistency value that could be obtained for any student was 0 if a student earned a "0" for each item. To summarize, higher values reflected more consistency between concurrent and retrospective reports for students, whereas lower values reflected less consistency between concurrent and retrospective reports. Consistency values were evaluated by student achievement level. Students' median mathematic score was used as a cut score to divide the 71 students into two sub-samples. That is, a median of their mathematics scores (M = 82) was used to differentiate the high-achieving students from the moderate-achieving students.

Consistency values were also evaluated by interviewer knowledge level. Interviewer knowledge level comprised three levels: high knowledge in mathematics (i.e., expert condition), low knowledge in mathematics (i.e., novice condition), and no information about knowledge (i.e., control condition).

Consistency values were also evaluated by item difficulty. Item difficulty comprised three levels: five easy items, five moderate items, and five difficult

items. The main interest here was to determine whether students' consistency between concurrent and retrospective verbal reports was affected by item difficulty. Therefore, the consistency values were summed for easy, moderate, and difficult items under each coding scheme. The highest consistency value that could be obtained for each student was 5 (across coding 1, 2, and 3) if a student earned a "1" for each item. The lowest consistency value that could be obtained for any student was 0 if a student earned a "0" for each item. For the last coding scheme, the highest consistency value that could be obtained for each student was 10 if a student earned a "2" for each item. The lowest consistency value that could be obtained for any student was 0 if a student earned a "0" for each item. Again, higher values reflected more consistency between concurrent and retrospective reports, whereas lower values reflected less consistency between concurrent and retrospective reports.

CHPATER III - RESULTS

This chapter presents results to the following research questions as outlined in the Methods section:

1. Are concurrent and retrospective verbal reports consistent across math test items?

2. Is the consistency of concurrent and retrospective reports influenced by student achievement level? Specifically, between high achieving and moderate achieving students?

3. Is the consistency of concurrent and retrospective reports influenced by interviewer knowledge level?

4. Is the consistency of concurrent and retrospective reports influenced by item difficulty (i.e., easy, moderate, vs. difficult items)?

It is important to note at the outset that the analysis included only 69 students instead of 71 because two students' verbal data were found to be corrupted. For example, one student mumbled throughout the interview, leading to indecipherable responses to the 15 mathematics items. The other student's verbal reports to the 15 mathematics items were not considered adequate because the conditions of the interview for this student were different from the conditions set up for the remainder of the students; namely, the interview of this student was interrupted by a fire alarm. The commotion of the fire alarm altered the concentration of the student and the interviewer expressed doubt regarding the quality of the students' responses. For the remainder of cases, the consistency

between concurrent and retrospective reports was evaluated according to four different criteria (coding schemes) of consistency as explained in the Methods section. An alpha level of .05 for all statistical tests was used in this study. The results are presented in the order of the research questions posed. For each research question, the results are presented from the most restrictive criterion of consistency to the least restrictive criterion of consistency (see Methods section, *Coding of Think-aloud Interviews).*

Consistency of Concurrent and Retrospective Reports: Descriptive Analysis

Descriptive analyses were conducted on the consistency of concurrent and retrospective verbal reports. First, the consistency between concurrent and retrospective reports for each item was investigated in light of four coding schemes (see Table 3). For example, as shown in Columns 3 and 4 of Table 3, when using the strictest coding scheme (coding scheme 1), 42 of 69 students, or 60.9%, exhibited perfect consistency in cognitive models between concurrent and retrospective reports for item 1; whereas 27 of 69 students, or 39.1%, displayed no consistency in models between concurrent and retrospective reports for item 1. For the second coding scheme, as shown in Columns 5 and 6 of Table 3, 26 of 69 (37.7%) students' verbal reports were consistent in the ability type of cognitive models displayed between concurrent and retrospective reports for item 1. That is, 26 students reflected cognitive models of similar ability levels for their concurrent and retrospective reports. Similarly, for the third coding scheme, as shown in Colum 7 and 8, 25 of 69 (36.2%) students' verbal reports were consistent in the cognitive models displayed between concurrent and retrospective reports for item

1. Namely, 25 students reflected cognitive models that matched the models developed by one of the test developers. The last two columns in Table 3 reflect a coding scheme designed to show students' complete, partial, or lack of consistency in model use. For example, for Item 1, 29 of 69 (42.0%) students' verbal reports demonstrated the use of some type of cognitive model between concurrent and retrospective reports. In other words, if a student's concurrent *and* retrospective verbal reports on item 1 reflected some type of cognitive model, their consistency was assigned a value of "2". However, if only one of a student's concurrent *or* retrospective reports displayed a cognitive model, their consistency was assigned value of "1" to show the presence of a partial use of models (according to Table 3, 31.9% of students demonstrated partial use of models). If a student's concurrent or retrospective report did not show any use of a model, their consistency was assigned a value of "0". This last coding (coding scheme 4) scheme reflects students' tendency to use any kind of cognitive model in their verbal reports.

From Table 3, we can also see that the proportion of students showing consistency on each item differs across the four coding schemes. Results were broken down by level of item difficulty for ease of interpretation. The first 5 items in Table 3 (items 1 to 5) are easy items, the next 5 items (items 6 to 10) are moderate items, and the last 5 items (items 11 to 15) are difficult items.

For easy items, the first coding scheme yielded a majority of students responding with consistent cognitive model use for all five items. For example, for item 1, 60.9% of students exhibited consistency in cognitive models between

their concurrent and retrospective reports. Moreover, for items 2 through 5, the percentage of students exhibiting consistency was greater than 50%. This was not the case, however, for the other coding schemes. Consider the easiest items in the second coding scheme. Although three of five items (item 3, 4 and 5) were associated with a proportion greater than 50% of students exhibiting consistency in their model use (e.g., 53.6% for item 3), two of the five items (item 1 and 2) were associated with a majority of students exhibiting *inconsistency* in their model use (e.g., 62.3% of students exhibited inconsistency for item 1). For the third coding scheme, a greater proportion of students showed consistent model use between concurrent and retrospective verbal reports. For example, the proportion of students showing consistent model use between concurrent and retrospective reports was over 50% for items 2, 3, 4, and 5. However, the proportion of students showing inconsistent model use was under 50% for item 1 at 36.2%. For the last coding scheme, the proportion of students who displayed some type of model use was dramatically greater than the proportion of students who displayed no model use. For example, a majority of students showed a fully consistent or partially consistent use of models (i.e., value of 1 or 2) between their concurrent and retrospective verbal reports for all five items.

Among moderate items, for the first coding scheme, a greater proportion of students displayed consistent model use between concurrent and retrospective verbal reports for item 7, 8, 9, and 10 (e.g., 59.4% for item 8), but not for item 6 (i.e., 49.4%). For the second coding scheme, three of five items (items 7, 9, and 10) were associated with a majority of students exhibiting consistency in their

model use (e.g., 58% for item 7). Nevertheless, two of five items (item 6 and 8) were associated with less than 50% of students exhibiting consistency in their model use. According to the third coding scheme, a greater proportion of students exhibited consistency between their concurrent and retrospective verbal reports for item 6, 7, 9 and 10 (e.g., 58% for item 9), but not for item 8 (e.g., 36.2%). For the last coding scheme, the proportion of students who displayed some kind of model use was again greater than the proportion of students who displayed no model use across the five moderate items (e.g., 91.3% when 10.1% and 81.2% are added for item 10).

As for the last five difficult items, for the first coding scheme, a greater proportion of students displayed consistency in their model use between concurrent and retrospective verbal reports for items 11, 12, 14, and 15 (e.g., 62.3% for item 11), but not for item 13 (i.e., 47.8%). For the second coding scheme, a majority of students exhibited consistency in their model use between concurrent and retrospective verbal reports for items 11, 12, 13, and 14 (e.g., 59.4% for item 12), but not for item 15 (i.e., 14.5%). A similar pattern was found for the third coding scheme; that is, a greater proportion of students displayed consistency in their model use for items 11, 12, 13, and 14 (e.g., 66.7% for item 14), but not for item 15 (i.e., 14.5%). For the last coding scheme, a majority of students displayed some kind of model use between concurrent and retrospective verbal reports across the five difficult items (e.g., 92.8% when 7.2% and 85.5% are added for item 14).

The proportions displayed in columns 3 and 4 of Table 3 need more elaboration as they might falsely suggest that students are highly consistent in specific model use. It is important to note that under coding scheme 1, students who displayed no model use in both concurrent and retrospective reports were classified as being *consistent* in their verbal reports. Thus, as shown in Table 4, for item 1, 18 of 69 (26.09%) students' verbal reports were classified as showing "no model" for both concurrent and retrospective verbal reports. The proportion of 26.09% in Table 3 is included in the overall proportion of 60.9% for item 1, coding level 1.

Second, the consistency of reports according to the four coding schemes across all the items was investigated. Shown in Table 5 are the central tendency, standard deviations and other supplementary descriptive statistical results. For the first coding scheme (i.e., if for a given item and a given student, the concurrent report and retrospective report showed an identical cognitive model classification, then a value of 1 was assigned for consistency, otherwise a zero), the consistency scores for this coding scheme were bi-modally distributed (see Figure 2). The, median was used as a central tendency measure of the consistency scores (M = 11.00, range of possible scores from 0 to 15) for the first coding scheme since the median was considered a better measure for this distribution (Gravetter & Wallnau, 2007, p.91). As shown in Table 5, this result indicates that 50% of students fell above and below 11 points for their consistency. For the second coding scheme, the consistency scores for this coding scheme were approximately normally distributed (see Figure 3, M = 7.28, SD = 4.25, range of possible scores

from 0 to 15). This result shows that, across all 15 items, students on average demonstrated consistency in terms of ability–level of cognitive models used for concurrent and retrospective reports for approximately 7 items out of possible 15 items. For the third coding scheme, the consistency scores for this coding scheme were also approximately normally distributed (Figure 4, $M = 7.94$, $SD = 3.65$, range of possible scores from 0 to 15 as well). This result implies that, across all 15 items, students on average displayed consistency in the cognitive models used (as outlined by a single test developer) between concurrent and retrospective reports for approximately 8 items out of a possible 15 items. For the last coding scheme, the consistency scores for this coding scheme were normally distributed (Figure 5, $M = 22.23$, $SD = 4.62$, range of possible scores from 0 to 30). This shows that, across all 15 items, students on average demonstrated consistency in cognitive models used in concurrent and retrospective reports for approximately 11 items out of a possible 15 items. This last coding scheme displayed the least restrictive method to measure consistency across students' verbal reports, reflecting the tendency to use a cognitive model or a partial cognitive model in solving the items.

Table 5 shows one way of summarizing consistency of model use between concurrent and retrospective reports. An alternative way of summarizing consistency involves examining Table 3 results and noting the items for which over half of students demonstrated consistency. In this way, for coding scheme 1, students were consistent in their models use for 13 out of 15 items (i.e., items 1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, and 15). For coding scheme 2, students were

consistent in their model use for 10 out of 15 items. For coding scheme 3, students were consistent in their model use for 12 out of 15 items. Finally, for coding scheme 4, most students displayed some kind of model in either their concurrent or retrospective reports for 15 out of 15 items.

Student Achievement

In order to evaluate the effect of student achievement on the consistency between concurrent and retrospective verbal reports, four independent sample *t*-tests were conducted individually on the four different codings of consistency scores. As mentioned previously, there were two levels of student achievement: higher-ability achievement students ($n = 36$, $M = 92.30$, $SD = 5.71$) and moderate-ability achievement students ($n = 33$, $M = 68.39$, $SD = 9.83$). The null hypothesis tested is that there is no significant difference between the two student achievement groups on the consistency between concurrent and retrospective reports across each of the four different codings of consistency scores. Shown in Table 6 are the means and standard deviations by student achievement for the four different coding schemes.

For the first coding scheme, the assumption of equal variance was tested with Levene's test of equal variances, and results showed that the assumption was met ($p > .05$). The independent samples *t*-test results indicated there was no significant effect of student achievement levels, $t(67) = .44$, $p = .66$, $d = .11$, on the consistency between concurrent and retrospective reports. Namely, higher-achieving students were not any more consistent in the cognitive models exhibited by their verbal reports ($M = 9.06$, $SD = 5.44$) than moderate-ability students ($M =$

8.48, *SD* = 5.22) using the strictest consistency coding scheme (coding scheme 1). For the second coding scheme, again the assumption of equal variance was met, and the results indicated a statistically significant effect of student achievement, $t(67) = 2.74$, $p < .05$, $d = .66$. That is, higher-achieving students' concurrent and retrospective reports (*M* = 8.56, *SD* = 3.97) reflected greater consistency in the ability level of their models than moderate-achieving students' concurrent and retrospective reports (*M* = 5.88, *SD* = 4.16). For the third coding scheme, the assumption of equal variance was also met, and the results indicted a statistically significant effect of student achievement, $t(67) = 3.33$, $p < .05$, $d = .80$; that is, higher-achieving students' verbal reports were more consistent (*M* = 9.25, *SD* = 3.17) than moderate-ability students' concurrent and retrospective reports (*M* = 6.52, *SD* =3.65) in terms of the cognitive models outlined by a single test developer . For the last coding scheme, the assumption of equal variance was not met ($p < .05$), and an adjusted independent samples *t*-test showed a significant effect of student achievement, $t(67) = 6.129$, $p = .000$, $d = 1.46$. Namely, higher-achieving students' verbal reports (M = 24.86, SD = 2.76) showed greater use of cognitive models, even partial use of cognitive models, than moderate-ability students (M = 19.36, SD = 4.55). These results will be evaluated in the Discussion section.

Interviewer Knowledge

Four separate, one-way analyses of variance (ANOVA) were conducted on the four codings of consistency to evaluate the effect of interviewer knowledge on the consistency between concurrent and retrospective reports. As mentioned

previously, the students were randomly assigned to one of three interviewer knowledge level conditions. The conditions included the interviewer identifying himself or herself as an expert in mathematics, a non-expert in mathematics, or as neither (control condition). The null hypothesis being tested is that the expertise of the interviewer does not influence the consistency of students' concurrent and retrospective reports. Shown in Table 7 are the means and standard deviations of consistency values across the three conditions of interviewer knowledge for the four different coding schemes.

The assumption of homogeneity of variance was met for all four analyses ($p < .05$), with a sample size of 23 for each group. For the first coding scheme of consistency, there was no effect of interviewer expertise, $F(2, 66) = .44$, $p = .644$, $\omega^2 = .01$, indicating that no matter whether the interviewer identified himself or herself as an expert, novice, or neither in mathematics, the consistency of the concurrent and retrospective reports did not change. The same pattern of results was found for the second coding scheme, where again there was no effect of interviewer expertise, $F(2, 66) = .44$, $p = .65$, $\omega^2 = .01$ on the consistency of students' concurrent and retrospective reports. In other words, it did not matter whether students were told the interviewer was an expert or novice in mathematics, or nothing at all, concurrent and retrospective reports remained unchanged in the ability-level of the cognitive models exhibited between concurrent and retrospective reports. The third coding scheme also showed that there was no effect of interviewer expertise, $F(2, 66) = .508$, $p = .604$, $\omega^2 = .02$. That is, the consistency between concurrent and retrospective verbal reports in

terms of the knowledge and skills each test developer considered reflective of success on solving problems was similar regardless of whether students were told the interviewer was an expert, novice, or nothing at all. Lastly, there was no effect of interviewer knowledge on values assigned using the fourth coding scheme, $F(2, 66) = 1.57$, $p = .216$, $\omega^2 = .05$. Revealing again that no matter how the interviewer identified himself or herself, the students showed similar patterns of cognitive model use between concurrent and retrospective reports. These results are discussed in the Discussion section.

Item Difficulty

The effect of item difficulty on the consistency of students' verbal reports was evaluated using four one-way, repeated-measure ANOVAs. The repeated-measure ANOVAs included item difficulty as a repeated variable as all the students ($N = 69$) were asked to finish all five easy items, five moderate items, as well as five difficult items. Item difficulty had three levels: easy items, moderate items, and difficult items. The dependent variable included the four different coding schemes of consistency values. The null hypothesis for this analysis is that there is no significant difference in the consistency of students' concurrent and verbal reports across levels of item difficulty. The mean and standard deviations for the four coding schemes of consistency values across different item difficulty conditions are shown in Table 8.

With the assumption of sphericity met for all four analyses ($p > .05$), there were no statistically significantly effects of item difficulty on the consistency of students' concurrent and retrospective reports. Specifically, for the first coding

scheme, there was no effect of item difficulty, $F(2, 66) = 1.24$, $p > .05$, $\omega^2 = .018$, indicating that students' consistency in the cognitive models exhibited in their concurrent and retrospective reports did not change as a function of solving easy, moderate, or difficult mathematics items. For the second coding scheme, again there was no effect of item difficulty, $F(2, 66) = .045$, $p = .956$, $\omega^2 = .018$, indicating that the ability-level of the cognitive models exhibited in students' verbal reports remained consistent across easy, moderate, and difficult items. For the third coding scheme, again there was no effect of item difficulty, $F(2, 66) = 1.29$, $p > .05$, $\omega^2 = .019$. This null result again indicated that students' concurrent and retrospective reports remained consistent in showing the knowledge and skills identified by a single test developer across easy, moderate, and difficult items. For the last coding scheme, there was also no effect of item difficulty, $F(2, 66) = .27$, $p = .77$, $\omega^2 = .004$; that is, students' concurrent and retrospective reports were found to demonstrate use of a cognitive model, even a partial model, irrespective of whether the items being solved were easy, moderate, or difficult. Implications of these results are discussed in the next section.

CHAPTER IV - DISCUSSION

The objective of this study was to evaluate and compare students' concurrent and retrospective verbal reports for consistency in their cognitive models of learning. To meet this objective, verbal data from student responses to 15 multiple-choice mathematics questions were used. The procedures used to collect verbal report data in this study were studied because the procedures used to collect verbal report data can influence the quality of the data and the validity of inferences made about students (Ericsson & Simon, 1993). It is expected that the process of evaluating and comparing the consistency of concurrent and retrospective verbal reports will help improve the information derived from protocol analysis and the procedures used to collect verbal reports in educational measurement studies.

Three independent variables were investigated in the present experimental study: students' achievement level, interviewer knowledge level, and item difficulty. These variables were investigated to evaluate their influence on the consistency of students' concurrent and retrospective verbal reports. The concurrent and retrospective verbal reports were compared for consistency in students' use of cognitive models on 15 mathematics items. Consistency in concurrent and retrospective reports was used as the dependent variable in all analyses.

Results for the consistency between concurrent and retrospective reports were presented. Statistical analysis of students' concurrent and retrospective verbal reports indicated that students were generally consistent in using the same

cognitive models during problem-solving, but this consistency varied with the coding schemes used to evaluate the verbal reports. The discussion of results is presented sequentially with each research question posed in present study.

1. *Are concurrent and retrospective verbal reports consistent across math test items?*

Results of the overall consistency between concurrent and retrospective reports were generally positive. A majority of students showed consistency in their use of cognitive models for most of the mathematic test items, but the proportion of students displaying consistency varied across different items, as well as different coding schemes. Tables 3 and 5 summarized the consistency of cognitive models between concurrent and retrospective reports across the four coding schemes. The four coding schemes ranged from the strictest form of consistency (coding scheme 1) to the most accommodating form of consistency (coding scheme 4). As listed in Table 3, for the first coding scheme, which also happened to be the strictest coding scheme, more than half of the students exhibited consistency in cognitive models between concurrent and retrospective reports for 13 out of 15 mathematics items. There were only two exceptions, item 6 and 13. For example, for item 6, 50.7% of students were *inconsistent* in the cognitive models used between concurrent and retrospective verbal reports.

Using the second coding scheme, a majority of students displayed consistency in their use of cognitive models for 10 out of 15 items. However, a majority of students displayed inconsistent use of cognitive models for 5 items

(see Table 3). The inconsistencies in use of cognitive models under this coding scheme suggested that students were using cognitive models of a different ability level between their concurrent and retrospective reports (see Table 3). For example, 62.3% of students exhibited inconsistency in cognitive model use between their concurrent and retrospective verbal reports for item 1.

For the third coding scheme, a majority of students was found to exhibit consistency in cognitive model use for 12 out of 15 items. However, more than 50% of students exhibited inconsistent use of cognitive models for 3 items (see Table 3). Under the third coding scheme, consistency was determined by whether students systematically displayed cognitive models outlined by one of the test developers (regardless of ability level of the model). In item 1, for instance, 63.8% of students were inconsistent in the cognitive models displayed in their concurrent and retrospective verbal reports. For the last coding scheme, a high proportion of students displayed using some type of cognitive model in either their concurrent or retrospective verbal report (see Table 3).

In light of the four coding schemes, a summary of consistency values between concurrent and retrospective results is presented next. Under coding scheme 1, students on average showed consistency between concurrent and retrospective verbal reports in the cognitive models used for approximately13 items out of 15 items (see Table 3). Under coding scheme 2, students on average displayed consistency in their cognitive models between concurrent and retrospective verbal reports for 10 items out of 15 items. For coding scheme 3, students on average displayed consistency in the cognitive models used between

concurrent and retrospective verbal reports for approximately 12 items out of 15 items. Under the last, and also the least strict, coding scheme, students on average displayed some form of cognitive model in either the concurrent or retrospective report for all 15 items.[4] These results suggest that consistency between concurrent and retrospective verbal reports was generally observed for items. As shown in Table 3, although the proportion of students displaying consistency varied across coding schemes, generally over 50% of students were found to be consistent across a majority of items.

As mentioned in the introduction to this paper, consistency is sought between concurrent and retrospective reports as retrospective reports are expected to confirm or clarify the contents and processes outlined in concurrent reports (Leighton & Gierl, 2007b; Taylor & Donnie; 2000). The fact that retrospective reports were found to be generally consistent with concurrent reports in this study confirms that when collected immediately after the task is completed, retrospective verbal reports can yield consistent information about the cognitive processes students are using to answer the item. In other words, the information obtained in retrospective reports is similar to concurrent verbal reports (Ericsson & Simon, 1993).

---

[4] There is an alternate method of evaluating the overall consistency of verbal reports across items. So far in the discussion, Table 3 has been used to evaluate the items for which a majority of students show consistency across concurrent and retrospective reports. However, as shown in Table 5 in the Results section, an arithmetic mean of the proportions across items could also be calculated. The values shown in Table 5 are slightly different than those reported based on Table 3. As shown in Table 5, students on average showed consistent cognitive models for approximately half of the test items under the second and third coding schemes, and on average displayed consistency for approximately 9 and 11 out of 15 test items for coding schemes 1 and 4, respectively.

As for the finding of inconsistent verbal reports, one possible implication that could be deduced from these results is that collecting retrospective verbal reports is not necessarily redundant. Although students' information processing in concurrent and retrospective reports is expected to match in terms of the cognitive models used for problem solving, findings from this study indicate that this is not always the case. That is, for some items, concurrent and retrospective verbal reports elicited different information processing from students' as reflected by their cognitive models. An inspection of the items by content and difficulty indicated no pattern for which items might be responsible for consistent or inconsistent model use across concurrent and retrospective reports. It is interesting to note, however, that item 15, one of the most difficult items, led to inconsistent model use across coding schemes 2 and 3. One potential cause for inconsistent verbal reports may be that students found item 15 too difficult and selectively reported their solutions during the retrospective interviews.

Another possible explanation for the lack of consistency may lie in the nature of the cognitive model used to categorize students' verbal reports. By taking a closer look at the cognitive models used in the present study, it is found that some of the cognitive models may be too complex. For example, by reviewing the cognitive models developed by both test developers for item 8 (see Appendix A and Appendix B), it is apparent that some cognitive models involved a number of integrated steps, where multiple knowledge and skills were required for problem solving. For some items, parallel skills were required to solve items. Because skills and knowledge were needed simultaneously, it is possible that

students omitted some knowledge and/or skills when they were verbalizing their thinking process, especially in the retrospective verbal reports. Moreover, in retrospective reports, students may be trying to rationalize their answer or generalize their thought processes instead of tracing their thinking processes as they occurred in the concurrent interview (Ericsson & Simon 1993). Although the complexity of cognitive models were not manipulated in the present study, the results suggest that tasks requiring cognitive models that can be conceptualized and implemented in a linear manner (without simultaneous knowledge and skills) could allow students to better verbalize their thinking processes. Hence, linear-based cognitive models may be preferred for eliciting verbal reports. This also corresponds to what Ericsson and Simon (1993, p. 10) suggested: "We need a model in order to interpret data that are to be used, in turn, to test the model. Under these circumstances, our data-interpretation model should be as simple as possible…"

The results of this study suggest that when verbal reports are used to reveal students' problem solving in educational measurement, both concurrent and retrospective verbal reports are often consistent. As outlined in the literature review, retrospective reports play an important role in confirming or clarifying the contents and processes outlined in concurrent reports (Leighton & Gierl, 2007b; Taylor and Dionne; 2000). Consequently, in order to gain a consistent and accurate view of students' thinking processes, both concurrent and retrospective reports should be collected to verify the information gathered.

Difference between different coding schemes

It is noted that the proportion of students displaying consistency varied across different coding schemes. The results across different coding schemes need to be scrutinized. As the second and third coding schemes were identified as less strict methods of coding for consistency compared to the first coding scheme, it was expected that under these coding schemes, a majority of students would show consistency in cognitive model use for more items than the first coding scheme. However, this was not the case, as more items elicited inconsistent cognitive model use for the second coding scheme than the first. For example, a majority of students exhibited consistency for 13 items for coding scheme one but only 10 items for coding scheme two. However, it is important to note that the first coding scheme contained matches (marked as consistencies) for students who did not use any models (see Method section, Table 2 and Table 4). Under the first coding scheme, the criterion for consistency was met when concurrent and retrospective reports showed an identical cognitive model classification, including those students who showed no model use between the concurrent and retrospective reports. However, in the second coding scheme, consistency was denoted when the concurrent and retrospective reports were categorized into a cognitive model of similar ability. Although the second coding scheme was less strict, in the sense that more combinations of model use could lead to a classification of consistency, this definition excluded students who did not employ a cognitive model in both concurrent and retrospective reports. Consequently, when the proportion of students who displayed “no model” for both concurrent and retrospective verbal

reports was large for an item, the proportion of students showing consistency of model use for the item was small in coding scheme 2 (see also coding scheme 3). For example, as is illustrated in Table 4, for item 1, 18 out of 69 students, or 26.1% of students' verbal reports were classified as displaying "no model" for both concurrent and retrospective reports. These students' reports were classified as consistent under the first coding scheme, but inconsistent under the second coding scheme. This discrepancy resulted in fewer students showing consistency in cognitive model use under coding scheme 2. That is, 60.9% of students displayed consistency in cognitive models for item 1 under coding scheme 1 (but this proportion also included those students *who consistently did not use a model*). Coding schemes 1, 2, 3, and 4 need to be interpreted cautiously as they reflect distinct ways of summarizing the consistency of concurrent and retrospective reports.

In conclusion, although a majority of students exhibited consistency for 13 items under coding scheme 1, a large proportion of this consistency came from students who consistently did not use one of the cognitive models used to classify verbal reports. In future studies, different coding schemes could be developed in order to categorize those students displaying "no model" in terms of whatever knowledge and skills they are using to solve items.

Upon analyzing the consistency results by item, it was noted that item 15 had a lower frequency of students showing consistency in coding scheme 2 and coding scheme 3 compared to other items. For item 15, only 14.5% of students displayed consistency using cognitive models in coding scheme 2, and 14.5%

using coding scheme 3. Once again, this appears to be the case because of the high proportion of students displaying “no model” for both concurrent and retrospective verbal reports for item 15 (see Table 4, 44.94%). An alternative explanation for this item eliciting inconsistent verbal reports is the difficulty of the item. Only 17 students out of 69 provided the correct answers to this item, that is, 52 students got this item wrong.

In short, the first coding scheme used in the present study, which was also the strictest coding scheme, led to finding more overall consistency between concurrent and retrospective verbal reports. However, the high proportions of consistency under this coding scheme came from a large number of students displaying no models across concurrent and retrospective reports. With most students not using a cognitive model, the first coding scheme generated higher estimate of the consistency between the two types verbal reports for the item. Coding scheme 2 and scheme 3 performed better in detecting the real consistency between concurrent and retrospective verbal reports. Under coding scheme 2, over 50% of students showed consistency on 10 items. Under coding scheme 3, over 50% of students showed consistency on 12 items. Under coding scheme 4, all items elicited use of some type of cognitive model for concurrent and retrospective reports.

2. *Is the consistency of concurrent and retrospective reports influenced by student achievement level?*

Under coding scheme 1, student achievement did not have an effect on the consistency of cognitive model use between concurrent and retrospective reports. That is, higher-achieving students were not any more consistent than moderate-achieving students in using cognitive models between concurrent and retrospective reports. For the remaining three coding schemes, however, student achievement showed an influence on consistency in cognitive model use between concurrent and retrospective reports. Higher achieving students displayed higher consistency in cognitive models than moderate-achieving students under coding scheme 2, 3, and 4. The results suggest that higher-achieving students may be more focused in using a common strategy to solve items across concurrent and retrospective reports, compared to moderate-achieving students. Further, higher-achieving students reflected greater consistency than moderate-achieving students in terms of using cognitive models of a specific ability level (coding scheme 2) and using cognitive models developed by a single test developer (coding scheme 3). Lastly, higher-achieving students showed more full cognitive models or partial cognitive models across concurrent and retrospective reports than moderate-achieving students. Therefore, the results indicate that higher-achieving students demonstrate greater consistencies in model-based problem solving than moderate achieving students between concurrent and retrospective verbal reports.

During retrospective reports, students only need to recall their thoughts about how they solved the task. In contrast, during concurrent verbal reports, students have to work on the problems *and* verbalise their thinking processes at the same time. It is possible that this additional cognitive load impacted moderate-

achieving students and interfered with their consistency in model use compared to higher-achieving students. Moreover, because higher-achieving students are expected to have more comprehensive understanding of the domain and possess more knowledge and skills than moderate-achieving students, this deeper understanding may lead to more consistent application of knowledge and skills. Therefore, higher-achieving students appear better able to follow a clear path of problem solving in concurrent reports and remember their processing in retrospective verbal reports. Moderate-achieving students, on the other hand, may be more focused on performing the task due to their limited knowledge and skills, and may not pay as much attention to the process of thinking aloud at the same time. For these students, there may be greater variability in accurately documenting how they have solved the task. For moderate-achieving students, reporting their thinking processes during the retrospective report may be easier as the cognitive load is lightened. Upon a closer inspection of the data, it was found that of the 33 moderate-achieving students, only 2 reported less model use in retrospective reports than in concurrent reports.

3. *Is the consistency of concurrent and retrospective reports influenced by interviewer knowledge level*?

In general, the expertise of the interviewer did not have an effect on the consistency of cognitive model use between concurrent and retrospective verbal reports. There were no significant differences in consistency of cognitive model use under all four coding schemes across the three conditions of interviewer knowledge. These results suggest that no matter how the interviewer identified

himself or herself to students, either as an expert in mathematics, a non-expert in mathematics, or as a neutral interviewer, the consistency between concurrent and retrospective reports in cognitive models was similar using a variety of coding schemes for classifying consistency. That is, in think-aloud studies, the consistency between concurrent and retrospective verbal reports did not change as a function of at least one interviewer characteristic. One possible explanation for this finding is that during the think-aloud interview, students' perceived feelings about the interviewer will accompany the students from the start of the interview until the end of the interview. Therefore, any effect the knowledge-level of the interviewer has on students should not differentially influence the consistency of reports. If the students are nervous or anxious about the existence of the interviewer, this feeling will last until the end of the interview (and influence both concurrent and retrospective reports). Conversely, if the students do not have feelings of anxiety, then the existence of the interviewer will not influence the students for the whole interview.

4. *Is the consistency of concurrent and retrospective reports affected by item difficulty*?

Item difficulty did not have a significant influence on the consistency of cognitive model use between concurrent and retrospective verbal reports. As item difficulty increased, the consistency scores remained similar. For example, for easy items, moderate items, and difficult items, the consistency scores for first coding scheme were 3.01, 2.96, and 2.81 (out of 5) respectively (see Table 8). That is, the students on average presented perfect consistency between concurrent

and retrospective verbal reports for approximately 3 items out of a possible 5 easy items, 3 items out of a possible 5 moderate items, and 3 items out of a possible 5 difficult items. Therefore, item difficulty was not a factor in influencing the consistency in cognitive model use between concurrent and retrospective verbal reports under any coding scheme of consistency.

This study did not find evidence to suggest item difficulty had an influence on the consistency of cognitive model use between concurrent and retrospective verbal reports. However, it has to be noted that item 15 was anomalous compared to the other four difficult items (see Table 3). A majority of students displayed consistency in solving item 15 under coding scheme one, but this large proportion of consistency came from a large proportion of students consistently not using any of the cognitive models outlined in the present study (44.9%). That is, 31 out of 69 students' verbal reports were classified as showing "no model" for both concurrent and retrospective verbal reports for item 15. It is possible that this item elicited knowledge and/or skills that were different from those expected by the test developers who designed the cognitive models (e.g., students used other solutions rather than those reflected in the models) or students did not master the knowledge or skill required to correctly respond to the item (e.g., students did not know how to solve the problem). Although the existence of this anomalous item did not influence the general results, that is, item difficulty had no impact on the consistency between concurrent and retrospective verbal reports, tasks used to elicit verbal reports should be selected carefully. After all,

verbal data can hardly provide useful evidence of students' thinking processing if students' cannot solve the problem.

Limitation of the study and future directions

Results from this study provide educational researchers and practitioners with some insights into the consistency of concurrent and retrospective reports in the domain of educational measurement. However, a limitation of this study is that quantitative analysis was primarily used to evaluate consistency of verbal reports using four predetermined cognitive models. The verbal reports that did not fit into one of these four models were classified as *no model* use. In other words, when student verbal reports were coded as "no model," we did not distinguish between students who did not know how to solve the problem from students who used other strategies such as test wise-ness to solve the problem. Therefore, additional qualitative analysis could have aided the analysis and should be used in future studies to compare the contents of students' concurrent and retrospective verbal reports. A qualitative analysis could provide more insights into the consistency of concurrent and retrospective reports when a variety of models are used by students.

Another limitation of this study is the use of a self-developed coding scheme for model use consistency. Previous studies have relied on the use of a coding protocol suggested by Bettman and Park (1979) (see also Ericsson & Simon, 1993). However, the coding protocol used by Bettman and Park (1979) has been used to evaluate consumer choice behaviour and not academic cognitive

processing. Further, the coding schemes outlined by Ericsson and Simon focus on identifying the information processing steps in the verbal report and not on the consistency of these steps across concurrent and retrospective reports. Coding schemes were important for this study as consistency of verbal reports will depend on how the verbal data are coded. Different consistency coding schemes led to different results despite using the same verbal data. For example, in the present study, when students' verbal reports were classified as "no model" for both concurrent and retrospective reports, they were classified as consistent under coding scheme 1, but inconsistent under coding scheme 2. However, there was a large proportion of students displaying "no model" in both concurrent and retrospective verbal reports (see Table 4): 26.09% of students for item 1, 26.09% for item 2, 14.49% for item 6, 23.19% for item 8, 21.74% for item 9, and 44.93% for item 15. Depending on how these "no model" classifications were coded (i.e., excluded from the sample or coded as either consistent or inconsistent), the results of the consistency between concurrent and retrospective reports changed. Therefore, in future studies, researchers need to decide on what kind of coding scheme better represents the research questions being asked and the contents of verbal reports being investigated.

A final concern rests with the generalizability of the current study. It is noted that the present study used items developed for use on achievement tests, specifically, a Grade 12 mathematics assessment. The thinking processes underlying these mathematics test items may often require a linear or logical processing procedure that leads to a single correct answer. It is necessary to

conduct similar studies in other content subject areas that require fewer constraints on students' responses in order to investigate whether similar results will be found. Future studies should use other content-subject domains such as critical reading, social studies, and science to further explore the consistency between concurrent and retrospective verbal reports.

In summary, the results of present study indicate that concurrent and retrospective verbal reports of problem solving on a class of mathematics test items are generally consistent. In addition, we found that higher-achieving students tended to provide more consistent verbal reports than moderate achieving students. Moreover, the consistency of the two types of verbal reports did not change as a function of interviewer knowledge and item difficulty. Overall, the study indicates that eliciting verbal reports from both concurrent and retrospective generally leads to consistent information about students' cognitive models in the domain of educational measurement.

Table 1

Coding Value with the Corresponding Cognitive Model Name

Model Name	Code Value
KM moderate model	1
KM high model	2
HR moderate model	3
HR high model	4
NO cognitive model	5

Table 2

Four Coding Schemes from Most to Least Restrictive

Classification of Model		Coding Scheme			
Concurrent Report	Retrospective Report	Procedure 1	Procedure 2	Procedure 3	Procedure 4
1	1	1	1	1	2
1	2	0	0	1	2
1	3	0	1	0	2
1	4	0	0	0	2
1	5	0	0	0	1
2	1	0	0	1	2
2	2	1	1	1	2
2	3	0	0	0	2
2	4	0	1	0	2
2	5	0	0	0	1
3	1	0	1	0	2
3	2	0	0	0	2
3	3	1	1	1	2
3	4	0	0	1	2
3	5	0	0	0	1
4	1	0	0	0	2
4	2	0	1	0	2
4	3	0	0	1	2
4	4	1	1	1	2
4	5	0	0	0	1
5	1	0	0	0	1
5	2	0	0	0	1
5	3	0	0	0	1
5	4	0	0	0	1
5	5	1	0	0	0

Table 3

Frequency and Portion of Students Showing Consistency or Inconsistency on Each Item (N = 69)

		Consistency Criteria							
		Coding 1		Coding 2		Coding 3		Coding 4	
Item	Match Index	Frequency	percentage	Frequency	percentage	Frequency	percentage	Frequency	percentage
Item 1	0	27	39.1%	43	62.3%	44	63.8%	18	26.1%
	1	42	60.9%	26	37.7%	25	36.2%	22	31.9%
	2							29	42.0%
Item 2	0	31	44.9%	48	69.6%	32	46.4%	18	26.1%
	1	38	55.1%	21	30.4%	37	53.6%	12	17.4%
	2							39	56.5%
Item 3	0	27	39.1%	32	46.4%	24	34.8%	7	10.1%
	1	42	60.9%	37	53.6%	45	65.2%	13	18.8%
	2							49	71.0%
Item 4	0	24	34.8%	25	36.2%	27	39.1%	5	7.2%
	1	45	65.2%	44	63.8%	42	60.9%	13	18.8%
	2							51	73.9%
Item 5	0	28	40.6%	30	43.5%	30	43.5%	5	7.2%
	1	41	59.4%	39	56.5%	39	56.5%	20	29.0%
	2							44	63.8%
Item 6	0	35	50.7%	40	58.0%	34	49.3%	10	14.5%
	1	34	49.3%	29	42.0%	35	50.7%	13	18.8%
	2							46	66.7%
Item 7	0	29	42.0%	29	42.0%	29	42.0%	6	8.7%
	1	40	58.0%	40	58.0%	40	58.0%	12	17.4%
	2							51	73.9%
Item 8	0	28	40.6%	42	60.9%	44	63.8%	16	23.2%
	1	41	59.4%	27	39.1%	25	36.2%	26	37.7%
	2							27	39.1%
Item 9	0	17	24.6%	32	46.4%	29	42.0%	15	21.7%
	1	52	75.4%	37	53.6%	40	58.0%	12	17.4%
	2							42	60.9%
Item 10	0	32	46.4%	33	47.8%	21	30.4%	6	8.7%
	1	37	53.6%	36	52.2%	48	69.6%	7	10.1%
	2							56	81.2%
Item 11	0	26	37.7%	32	46.4%	30	43.5%	9	13.0%
	1	43	62.3%	37	53.6%	39	56.5%	12	17.4%
	2							48	69.6%
Item 12	0	29	42.0%	28	40.6%	33	47.8%	8	11.6%
	1	40	58.0%	41	59.4%	36	52.2%	13	18.8%
	2							48	69.6%
Item 13	0	36	52.2%	29	42.0%	28	40.6%	2	2.9%
	1	33	47.8%	40	58.0%	41	59.4%	9	13.0%
	2							58	84.1%
Item 14	0	31	44.9%	31	44.9%	23	33.3%	5	7.2%
	1	38	55.1%	38	55.1%	46	66.7%	5	7.2%
	2							59	85.5%
Item 15	0	29	42.0%	59	85.5%	59	85.5%	31	44.9%
	1	40	58.0%	10	14.5%	10	14.5%	25	36.2%
	2							13	18.8%

Table 4

Frequency and Percentage Students Displaying "No Model" for both Concurrent and Retrospective Verbal Reports

Item	Frequency	Percentage
item 1	18	26.09%
item 2	18	26.09%
item 3	7	10.14%
item 4	5	7.25%
item 5	5	7.25%
item 6	10	14.49%
item 7	6	8.70%
item 8	16	23.19%
item 9	15	21.74%
item 10	6	8.70%
item 11	9	13.04%
item 12	8	11.59%
item 13	2	2.90%
item 14	5	7.25%
item 15	31	44.93%

Table 5

Central tendency and Standard Deviations of Four Consistency Values for All Students

Coding scheme	N	M	SD	Minimum	Maximum
1	69	11.00	5.31	1	15
2	69	7.28	4.25	0	15
3	69	7.94	3.65	0	15
4	69	22.23	4.62	8	30

Note. In coding scheme 1, the distribution was a bimodal distribution. The central tendency used for this coding scheme was median. Mean was used since the distribution was close to normal distribution for coding scheme 2, 3, and 4.

Table 6

Means and Standard Deviations of Four Consistency Values for Student Achievement

	Student Achievement						
	High (*n*=36)		Low (*n*=33)				
Coding Scheme	M	SD	M	SD	P value	Total M	Total SD
1	9.06	5.44	8.48	5.22	.659	8.77	5.33
2	8.56	3.97	5.88	4.16	.008	7.22	4.06
3	9.25	3.17	6.52	3.65	.001	7.88	3.41
4	24.86	2.76	19.36	4.55	.000	22.11	3.65

Note. A Levene's test was conducted for each coding scheme. None of the results is significant ($p > .05$) for the coding scheme 1, 2, and 3. Therefore, we do not reject the null hypothesis of equal variance. To wit, the results indicate that there is no evidence to show that the variance is not equal between the two groups for the coding scheme 1, 2, and 3. The null hypothesis of equal variance was rejected for last coding scheme ($p < .05$). That is, the result indicates that there is evidence to show that the variance is not equal between the two groups for the last coding scheme.

Table 7

Means and Standard Deviations of Four Consistency Values for Interviewer Knowledge

	Interviewer Knowledge							
	Control (n=23)		Novice (n=23)		Expert (n=23)			
Coding Scheme	M	SD	M	SD	M	SD	Total M	Total SD
1	8.87	5.39	8.00	5.21	9.48	5.44	8.78	5.31
2	7.91	4.25	6.74	4.53	7.17	4.05	7.28	4.25
3	8.52	3.48	7.43	3.88	7.87	3.67	7.94	3.65
4	23.43	3.64	22.22	4.73	21.04	5.22	22.23	4.62

Note. A Levene's test was conducted on the variance for each sample. The results indicated that the variances could be considered statistically equivalent.

Note. No statistically difference was found for any of the interviewer knowledge effect on the consistency between concurrent and retrospective verbal reports reflecting cognitive models.

Table 8

Means and Standard Deviations of Four Consistency Values for Item Difficulty

	Item Difficulty							
	Easy (*N*=69)		Moderate (*N*=69)		Difficult (*N*=69)			
Coding Scheme	M	SD	M	SD	M	SD	Total M	Total SD
1	3.01	1.89	2.96	1.74	2.81	2.00	2.93	1.88
2	2.42	1.60	2.45	1.65	2.41	1.5	2.43	1.58
3	2.72	1.42	2.72	1.51	2.49	1.43	2.65	1.46
4	7.30	1.83	7.45	2.08	7.48	1.97	7.41	1.96

Note. No statistically difference was found for any of the item difficulty effect on the consistency between concurrent and retrospective verbal reports reflecting cognitive models.

Knowledge Level of the interviewer	Novice	
	Expert	
	Control	

Figure 1. Design of Between-subject Variables

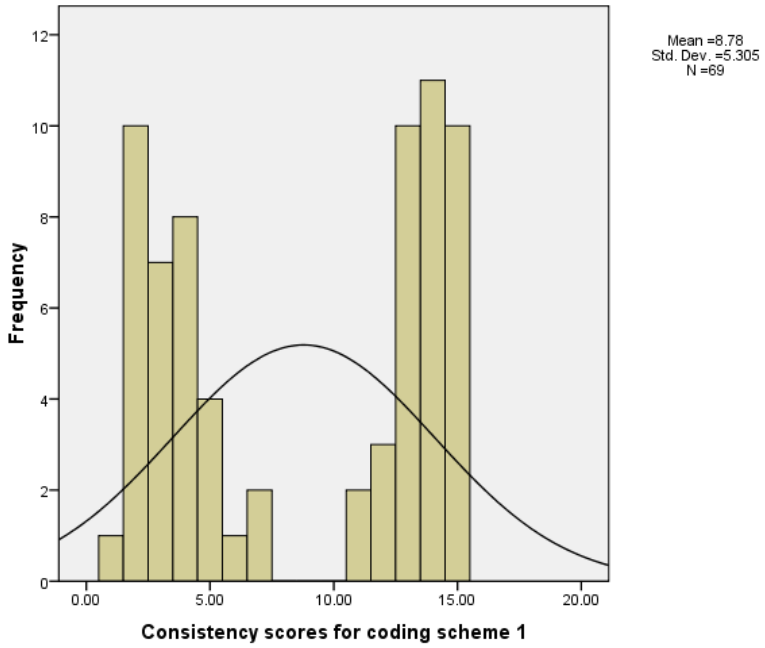


Figure 2. Consistency Scores of Students across All the Items for Coding Scheme 1

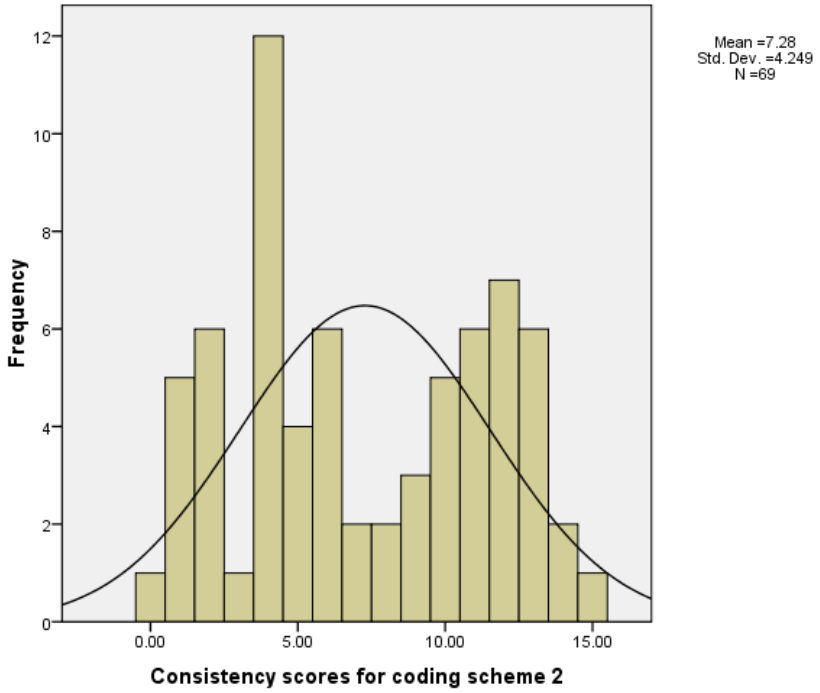


Figure 3. Consistency Scores of Students across All the Items for Coding Scheme 2

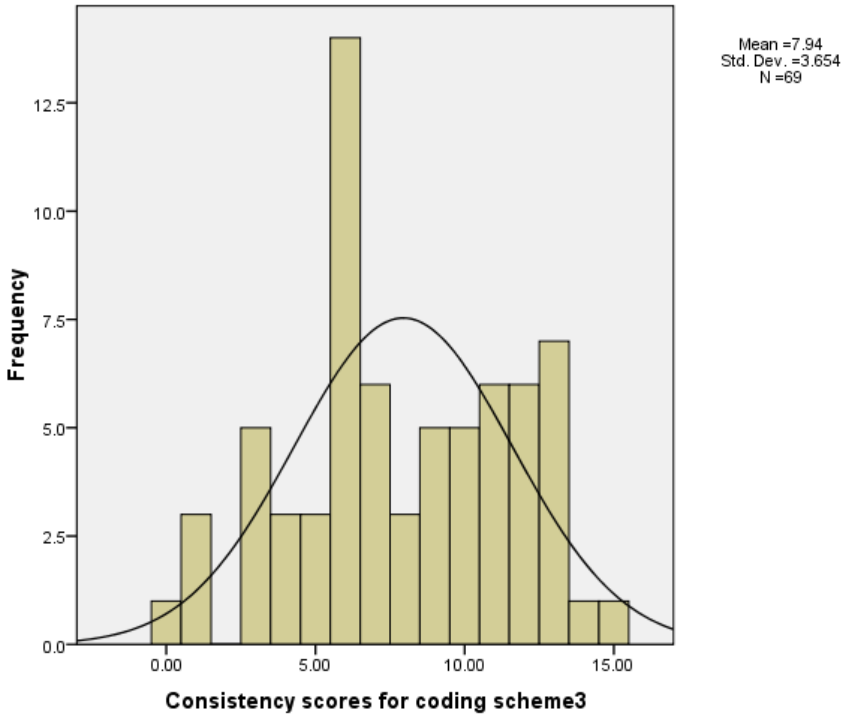


Figure 4. Consistency Scores of Students across All the Items for Coding Scheme 3

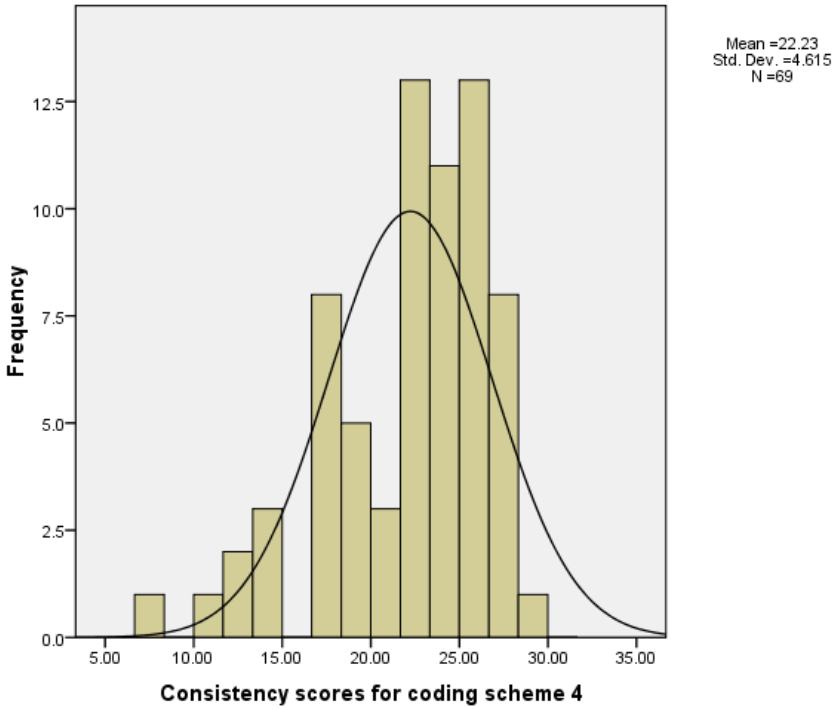


Figure 5. Consistency Scores of Students across All the Items for Coding Scheme 4

References

Afflerbach, P., & Johnson, P. (1984). On the use of verbal reports in reading research. *Journal of Reading Behavior, 16,* 307-322.

Alberta Education (2010). *Provincial Achievement Test: General Information Bulletin*. Edmonton, AB: Alberta Education.

Bettman, J, R., & Park, C.W. (1997). *Implications of a constructive view of choice for analysis of protocol data: A coding scheme for elements of choice processes* (Working Paper No. 75). Los Angeles: Center for Marketing Studies, University of California.

Camps, J. (2003). Concurrent and retrospective verbal reports as tools to better understand the role of attention in second language tasks. *International Journal of Applied Linguistics, 13*(2), 201–221.

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

College Board (2010), *Scholastic Aptitude Test*. Retrieved on November 1, 2010, from http://sat.collegeboard.com/about-tests.

Council of Ministers of Education, Canada (CMEC) (2007), *PCAP-13: Report on the Assessment of 13 Year Olds in Reading, Mathematics and Science.* Ottawa, CA: Council of Ministers of Education, Canada

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*(3), 380-396.

Embretson, S. E. (1999). Cognitive psychology applied to testing. In F. T. Durso, R. S. Nickerson, R. W., Schvaneveldt, S. T. Dumais, D. S. Lindsay, & M. T. H. Chi (Eds.), *Handbook of Applied Cognition* (pp. 629-660). New York: Wiley.

Embretson, S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38,* 343–368.

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

Ericsson, K. A.,&Simon, H. A. (1980).Verbal reports as data. *Psychological Review, 87,* 215–251.

Educational Testing Service (ETS) (2010), *Graduate Records Examination*. Retrieved on November 1, 2010, from http://www.ets.org/gre.

Frederiksen, N., Glaser, R., Lesgold, A., & Shafto, M. G. (Eds.). (1990). *Diagnostic monitoring of skill and knowledge acquisition.* New Jersey: Lawrence Erlbaum Associates.

Gierl, M. J. (1997). Comparing the cognitive representations of test developers and examinees on a mathematics achievement test using Bloom's taxonomy. *Journal of Educational Research, 97,* 26-32.

Gierl, M. J., Cui, Y., & Hunka, S. (2008). Using connectionist models to evaluate examinees' response patterns on tests. *Journal of Modern Applied Statistical Methods, 7*(1), 234-245.

Gravetter, F. J & Wallnau, L. B (2007). *Statistics for the Behavioral Sciences*. Australia; Belmont, CA: Thomson/Wadsworth, c2007.

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

Kane, M. T. (2006). Validation. In R. L. Brennan (ed.), *Educational Measurment 4th edition*. (pp. 17-64). Westport: American Council on Education and Praeger Publishers.

Koretz, D., & Hamilton, L. S. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531-578). Washington, DC: American Council on Education.

Kuusela, H., & Paul, P. (2000). A comparison of concurrent and retrospective verbal protocol analysis. *American Journal of Psychology, 113*(3), 387-404.

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

Leighton, J. P. (2009). *Two Types of Think Aloud Interviews for Educational Measurement: Protocol and Verbal Analysis.* Paper presented for symposium How to Build a Cognitive Model for Educational Assessments at the 2009 annual meeting of the National Council on Measurement in Education (NCME), April 14-16, San Diego, CA.

Leighton, J. P. (2010, October). *Using verbal reports as a source of data in validity studies*. Colloquium given at Educational Testing Service (ETS) in Princeton, New Jersey, USA.

Leighton, J.P. (2011, April). *Item difficulty and interviewer knowledge effects on the accuracy and consistency of examinee response processes in verbal* reports. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA, USA.

Leighton, J. P., & Gierl, M. J. (2007a). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice, 26,* 3–16.

Leighton, J. P., & Gierl, M. J. (2007b). Verbal reports as data for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications.* (pp. 146–172). Cambridge, UK: Cambridge University Press.

Leighton, J. P., & Gierl, M. J. (in press). *The learning sciences in educational assessment*. Cambridge, MA: Cambridge University Press.

Leighton, J. P., *&* Gokiert, R. *(*2005*). The cognitive effects of test item features: Identifying construct irrelevant variance and informing item generation.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Leighton, J. P., Gierl, M. J., & Hunka, S. (2004). The attribute hierarchy model: An approach for integrating cognitive theory with assessment practice. *Journal of Educational Measurement, 41,* 205-236

Leighton, J., P., Heffernan, C., Cor, M. K., Gokiert, R. J., & Cui, Y. (2008). *An Experimental Test of Student Verbal Reports and Expert Teacher*

Evaluations as a Source of Validity Evidence for Test Development. Annual Meeting of the American Educational Research Association, New York, NY.

Lohman, D. F. (2000). Complex information processing and intelligence. In R.J. Sternberg (Ed.), *Handbook of intelligence* (pp. 285–340). NY: Cambridge University Press.

Mislevy, R. J (2006). Cognitive Psychology and Educational Assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257–306). Washington, DC: American Council on Education.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33,* 379-416.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Nichols, P. (1994). A framework of developing cognitively diagnostic assessments. *Review of Educational Research, 64*, 575–603.

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (Eds.). (1995). *Cognitively diagnostic assessment.* Hillsdale, NJ: Erlbaum.

Nitko, A. J., Brookhart, S. M. (2007). *Educational Assessment of Students* (5th Edition). Upper Saddle River. NJ: Pearson Education.

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.

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

Organization for Economic Cooperation and Development (OECD) (2007), *PISA 2006 Science Competencies for Tomorrow's World Volume 1: Analysis*. Paris: OECD.

O'Neil, H. F. & Abedi, J. (1996). Reliability and Validity of a State Metacognitive Inventory: Potential for Alternative Assessment. *The Journal of Educational Research, 89*(4), 234-245.

Payne, J. W., Braunstein, M. L., & Carroll, J. S. (1978). Exploring predecisional behavior: An alternative approach to decision research. *Organizational Behavior and Human Performance*, 22, 17–44.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practices. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of Research in Education* (pp. 307-353). Washington, DC: American Educational Research Association.

Poggio, A., Clayton, D. B., Glasnapp, D., Poggio, J., Haack, P., & Thomas, J. *(*2005*). Revisiting the item format question: Can the multiple choice format meet the demand for monitoring higher-order skills*? Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal , Canada.

Popham, M. J. (2000). *Modern educational measurement: Practical guidelines for educational leaders* (3rd Edition). Needham, MA: Allyn and Bacon.

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading.* Hillsdale, NJ: Erlbaum.

Rogers, W. T., & Yang, P. *(*1996*).* Test-wiseness: Its nature and application. *European Journal of Psychological Assessment, 12,* 247–259.

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory & Cognition, 17,* 759–769.

Sawyer, T. P., Jr., & Hollis-Sawyer, L. A. (2005). A path-analytic assessment of different stress coping models in predicting stereotype threat perceptions, test anxiety reactions, and cognitive test performance. *International Journal of Testing, 5*(3), 225-246.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263-331). New York: American Council on Education, Macmillian.

Spielberger, C. D., Gonzalez, H. P., Taylor, C. J., Anton, W. D., Algaze, B., Ross, G. R., & Westberry, L. G. (1980). *Preliminary profession manual for the Test Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Standards for Educational and Psychological Testing. (1999). Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

Taylor, K. L., & Dionne, J-P. (2000). Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology, 98*, 413-425.

United States Department of Education (USDE) (2004). *Testing for results: Helping families, schools, and communities understand and improve student achievement*. In NCLB(Stronger accountability) (chap. 1). Retrieved February 15, 2006, from http://www.ed.gov/nclb/accountability/ayp/testingforresults.html.

Van den Haak, M.J., De Jong, M.D.T., & Schellens, P.J. (2003). Retrospective versus concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour & Information Technology,* 22, 339–351.

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

Van Gog, T., Paas, F., & Van Merriënboer, 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 Someren, M. W., Bamard, Y. F., & Sandberg, J. A. C. (1994). *The think aloud method: A practical guide to modeling cognitive processes.* London: Academic Press.

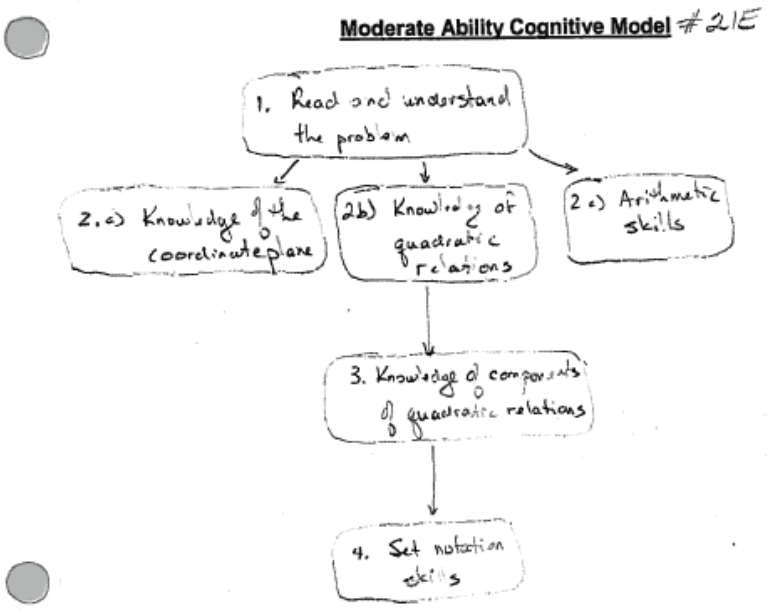
Wilson, T. D. (1994). The proper protocol: Validity and completeness of verbal reports. *Psychological Science, 5,* 249–252.

Appendices

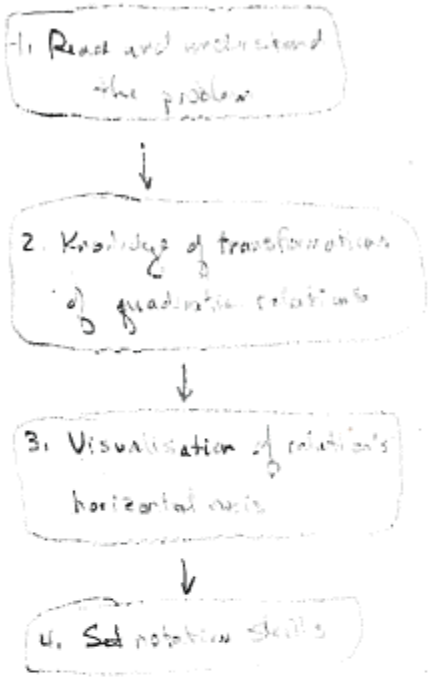
Appendix A – Cognitive models (flowchart) of the knowledge and skills that moderate ability and high ability students would use to solve each item correctly, developed by K.M.

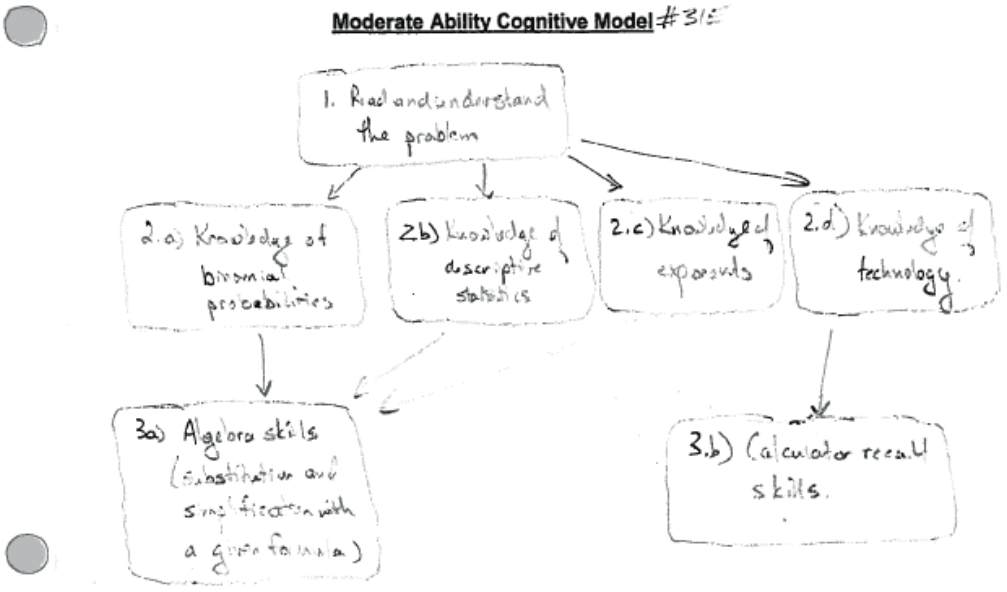
Appendix B – Cognitive models (flowchart) of the knowledge and skills that moderate ability and high ability students would use to solve each item correctly, developed by H.R.

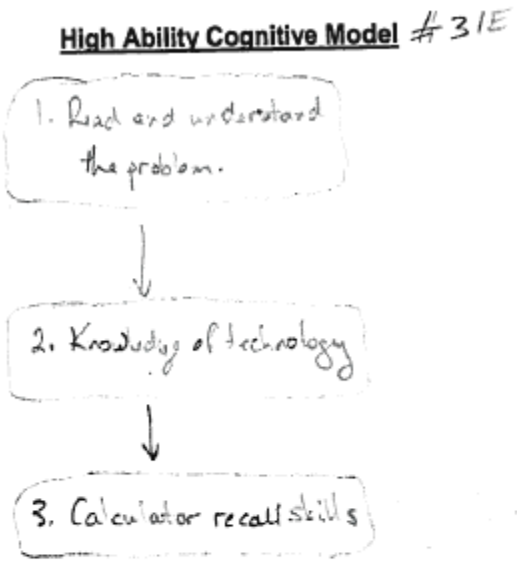
Appendix A – Cognitive models (flowchart) of the knowledge and skills that moderate ability and high ability students would use to solve each item correctly, developed by K.M.



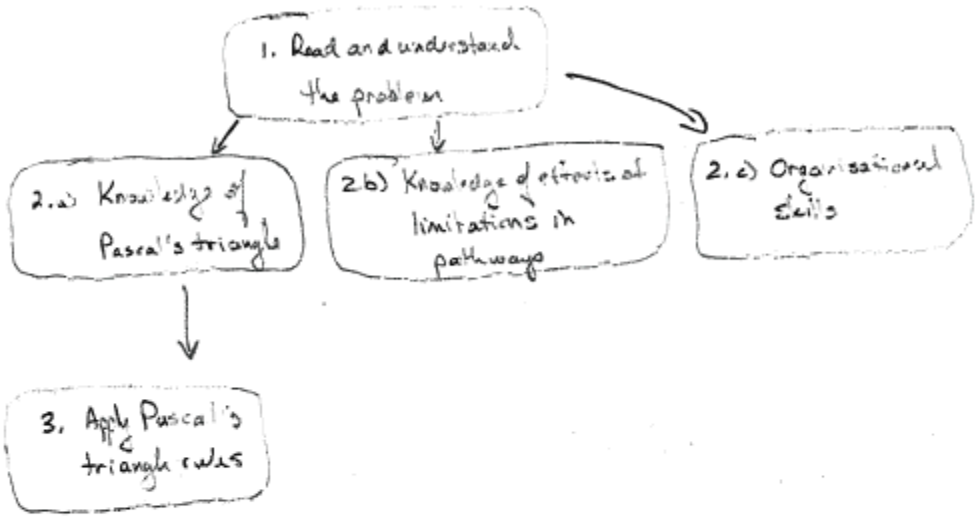
High Ability Cognitive Model #21E




Moderate Ability Cognitive Model #315
1. Read and understand the problem
2.a) Knowledge of binomial probabilities
2b) Knowledge of descriptive statistics
2.c) Knowledge of exponents
2.d) Knowledge of technology.
3a) Algebra skills (substitution and simplification with a given formula)
3.b) Calculator recall skills.


High Ability Cognitive Model #31E
1. Read and understand the problem.
2. Knowledge of technology
3. Calculator recall skills

Moderate Ability Cognitive Model #275

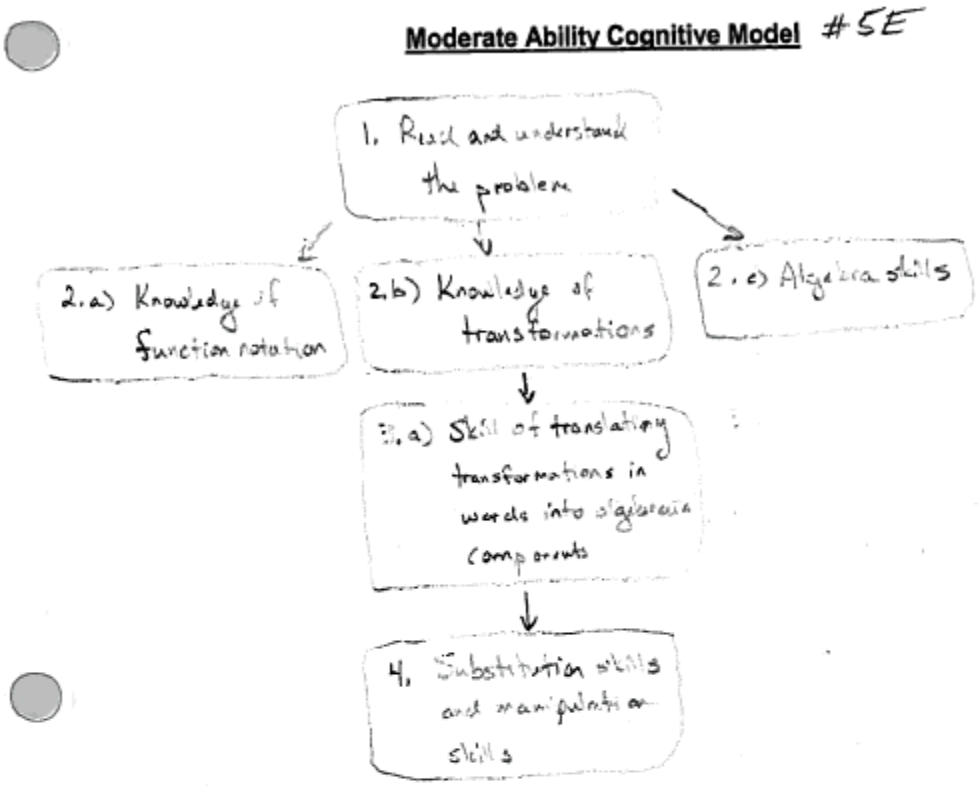


Notes: 2c) - one major reason that students get these pathway questions wrong is the inability to organise their working so that they keep the pattern moving logically... Conscientious examinees will often check these over 2 or 3 times!

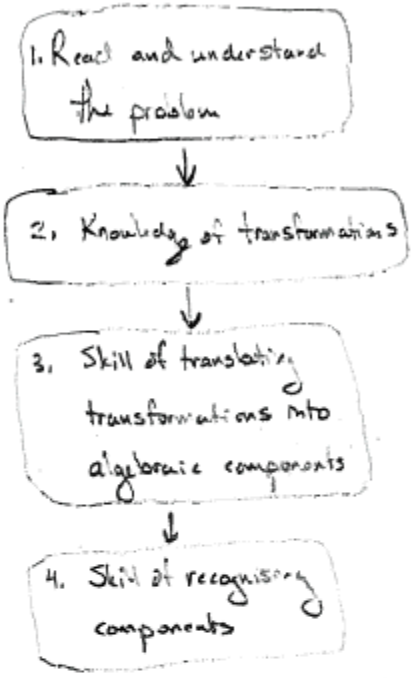
High Ability Cognitive Model # 27E

Same as moderate.

Notes: A higher ability student may use a little ~~tutwiseness~~ here. Choices C and D are too high (recognisably too high). The student will note that only A & B are related by a factor of 2, and, when they see the final square on the pathway, they will choose B.

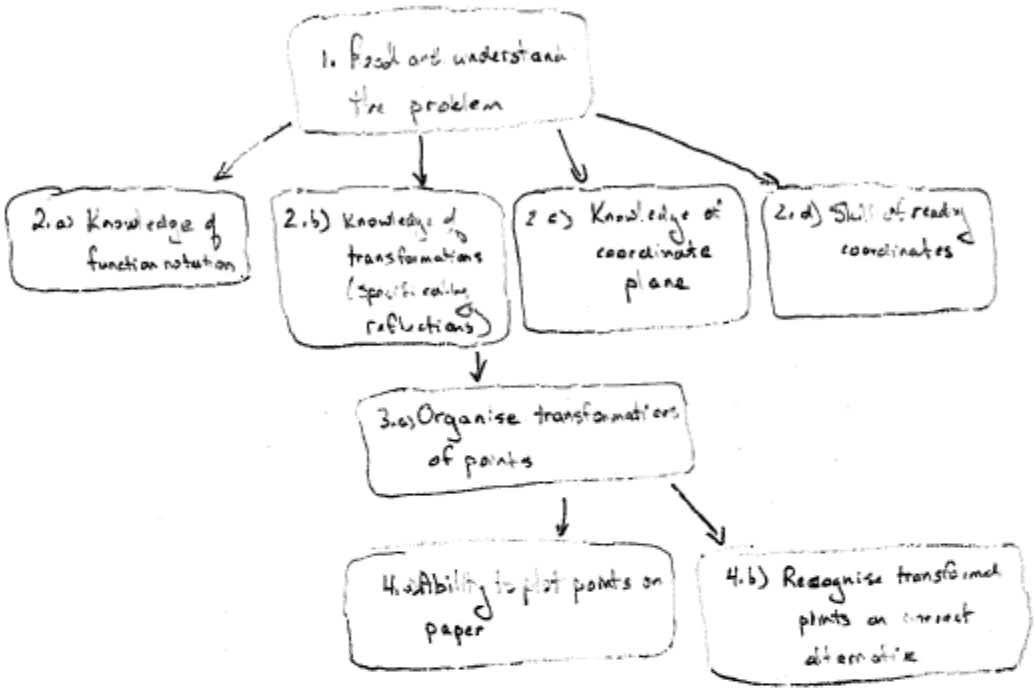

Moderate Ability Cognitive Model #5E
1. Read and understand the problem
2.a) Knowledge of function notation
2.b) Knowledge of transformations
2.c) Algebra skills
3.a) Skill of translating transformations in words into algebraic components
4. Substitution skills and manipulation skills

High Ability Cognitive Model #5E

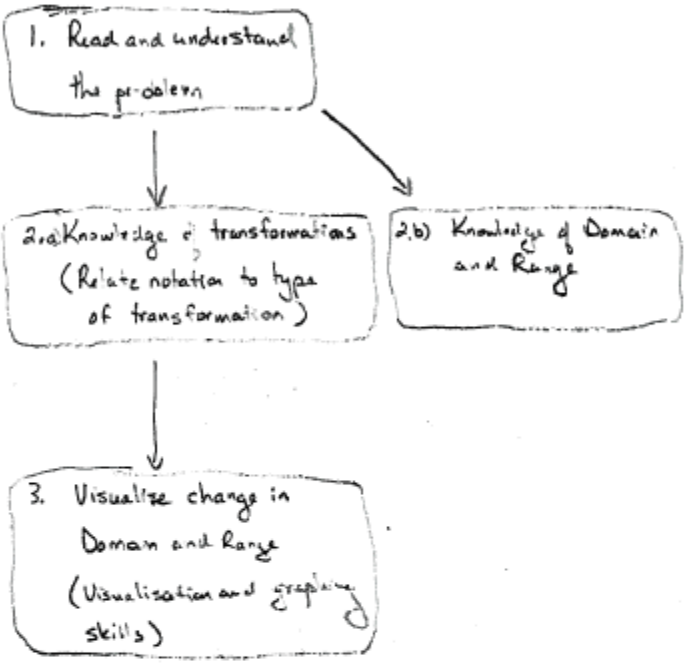


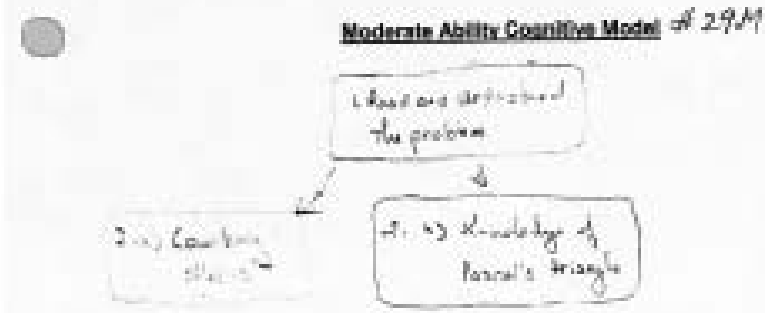
Notes: Since this is a multiple-choice item, high ability students will likely read the question, then discriminate the alternatives. They will see that the reflection and the vertical translation are already accounted for, and they will concern themselves ONLY with the two remaining transformations.

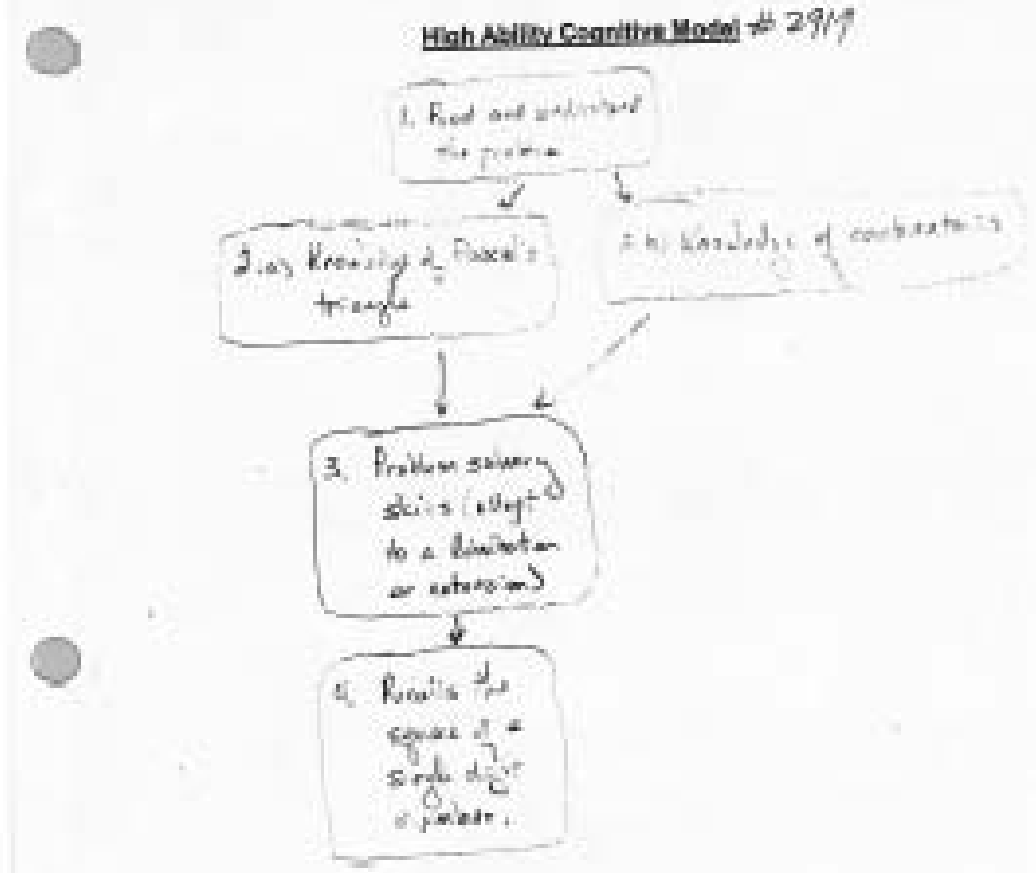
Moderate Ability Cognitive Model #4E

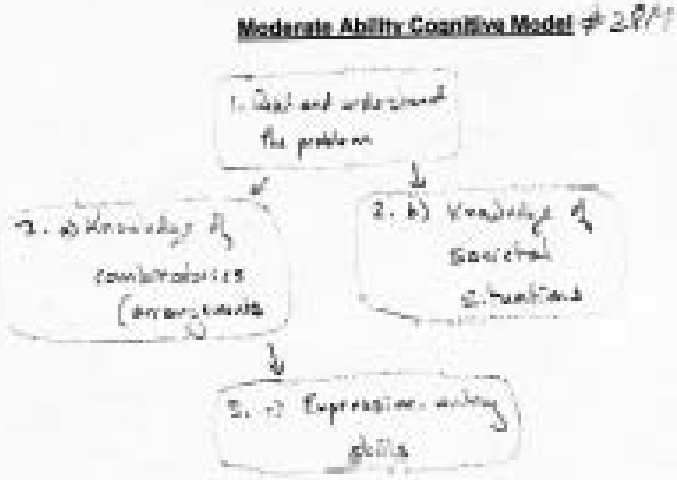


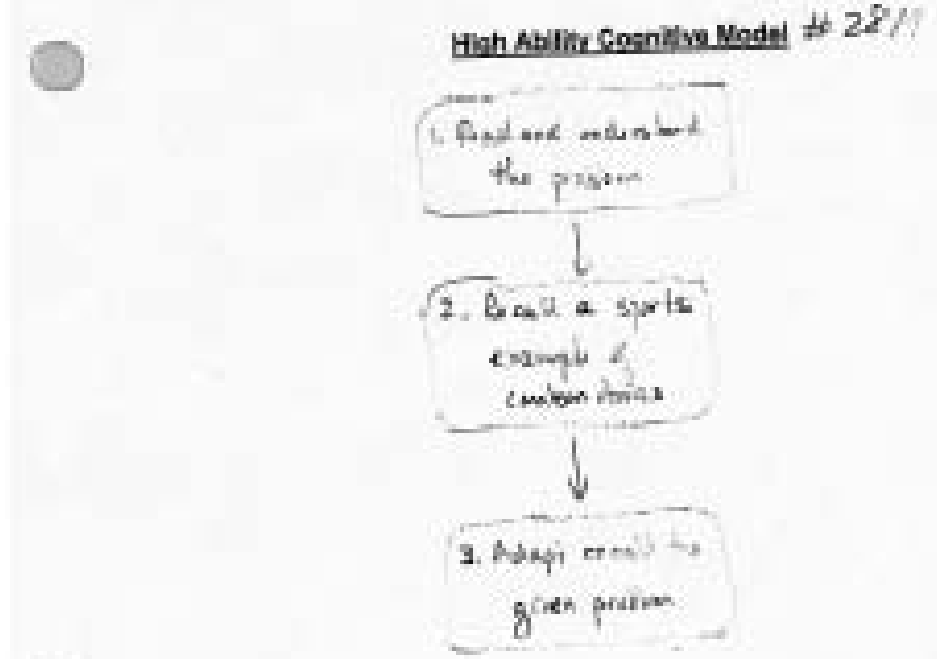
High Ability Cognitive Model #4

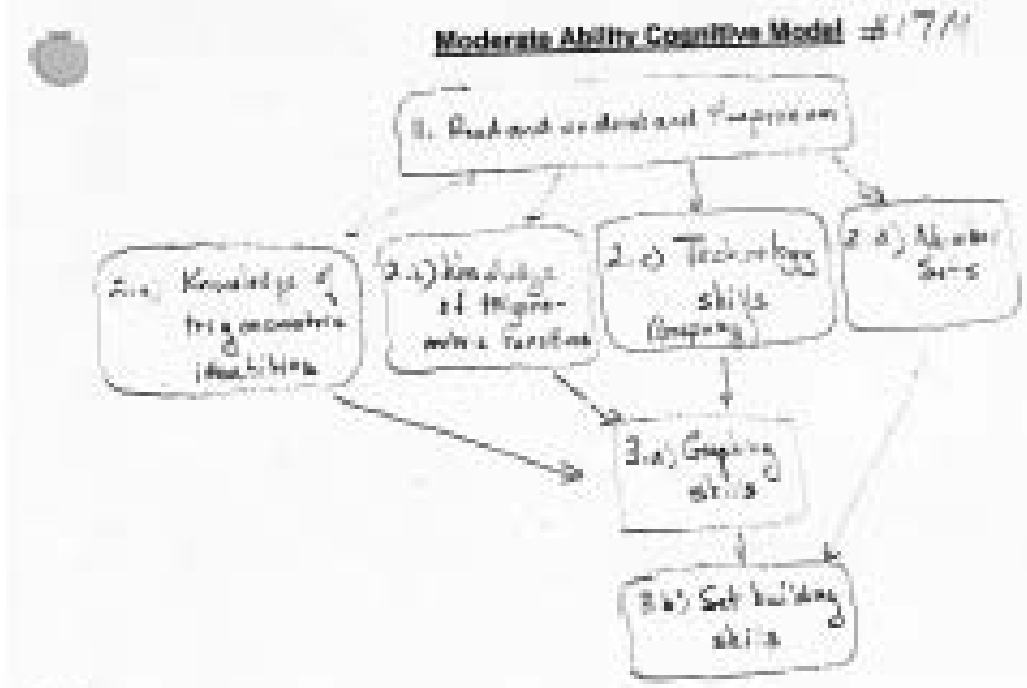


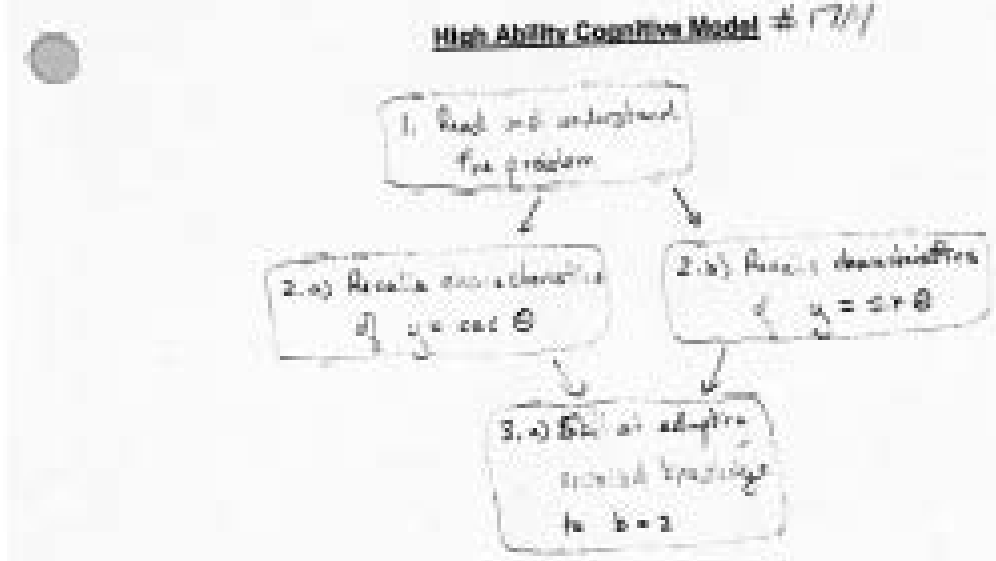

Moderate Ability Cognitive Model of 29M
Pascal's triangle

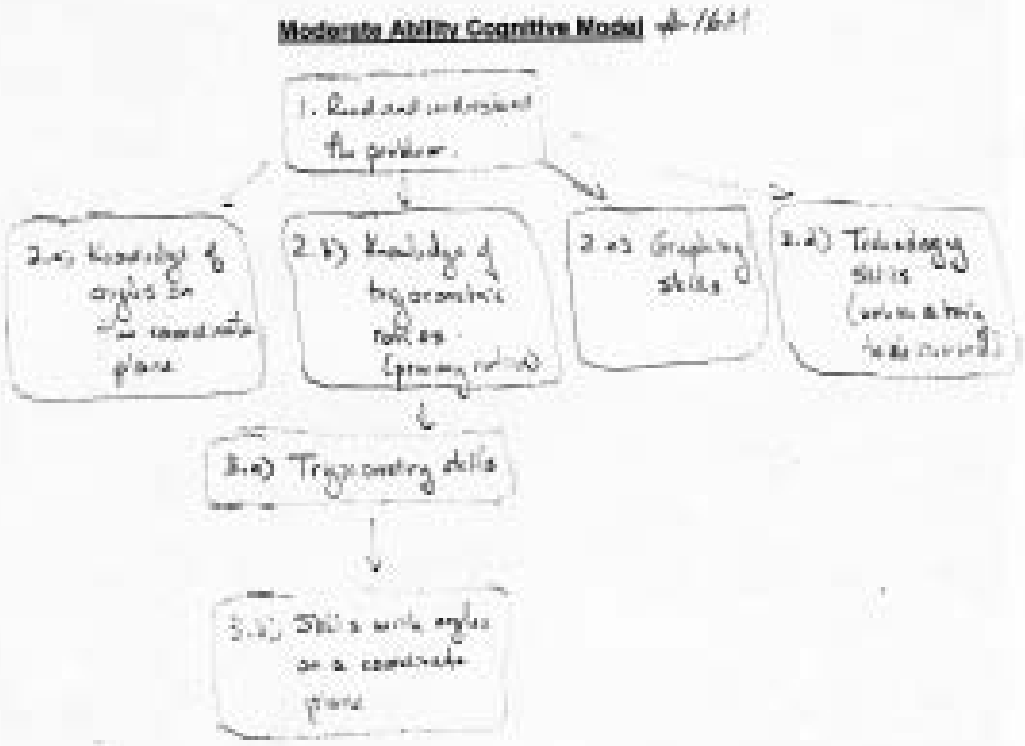

High Ability Cognitive Model #2917
1. Read and understand the problem
2. Use knowledge of Pascal's triangle
3. Problem solver skills (adapt to a limitation or extension)
4. Recalls the square of a single digit or number.

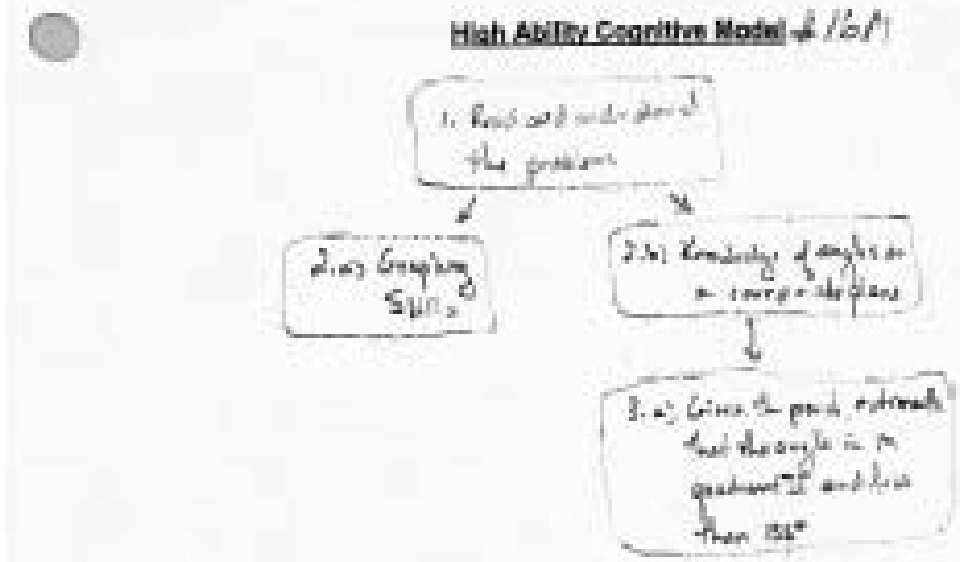

Moderate Ability Cognitive Model # 2
1. Read and understand the problem
2. a) Knowledge of computations
2. b) Knowledge of
3. Expressive writing skills

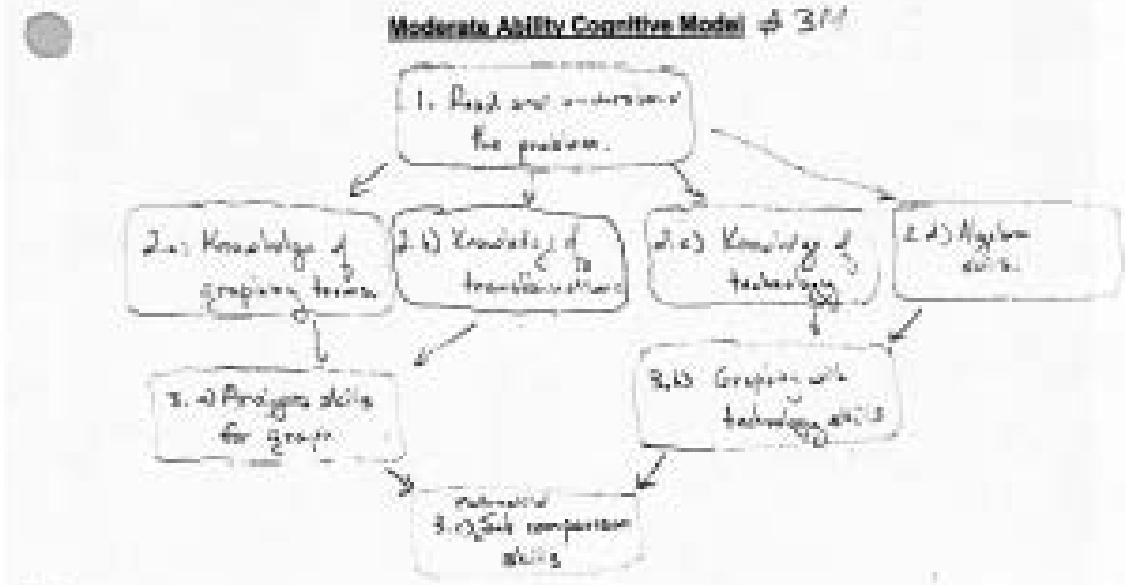

High Ability Cognitive Model #22
1. Read and understand the problem
2. Recall a specific example of
3.

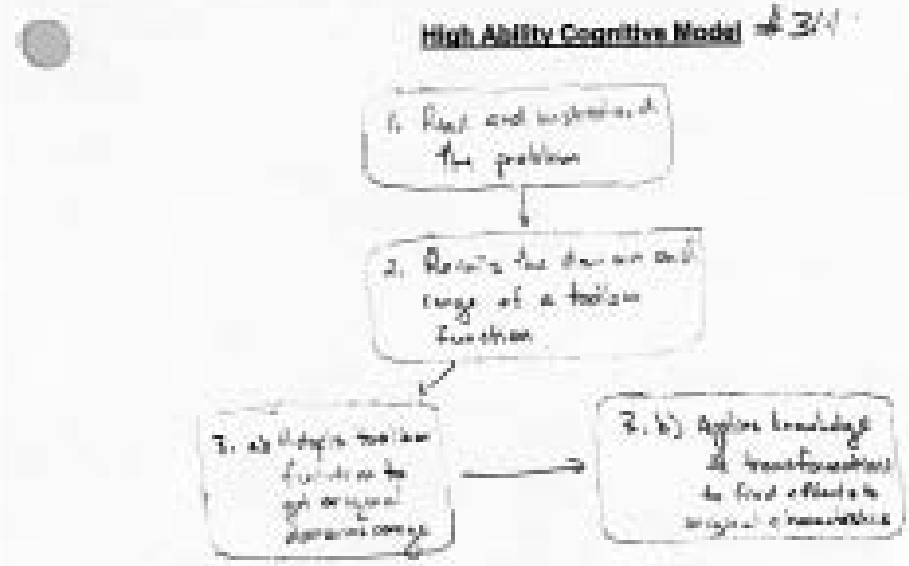

Moderate Ability Cognitive Model
1. Background academic experiences
2.a) Knowledge of trigonometric identities
2.b) Knowledge of trigonometric functions
2.c) Technology skills (Graphing)
2.d) Algebra Skills
3.a) Graphing skills
3.b) Self-building skills

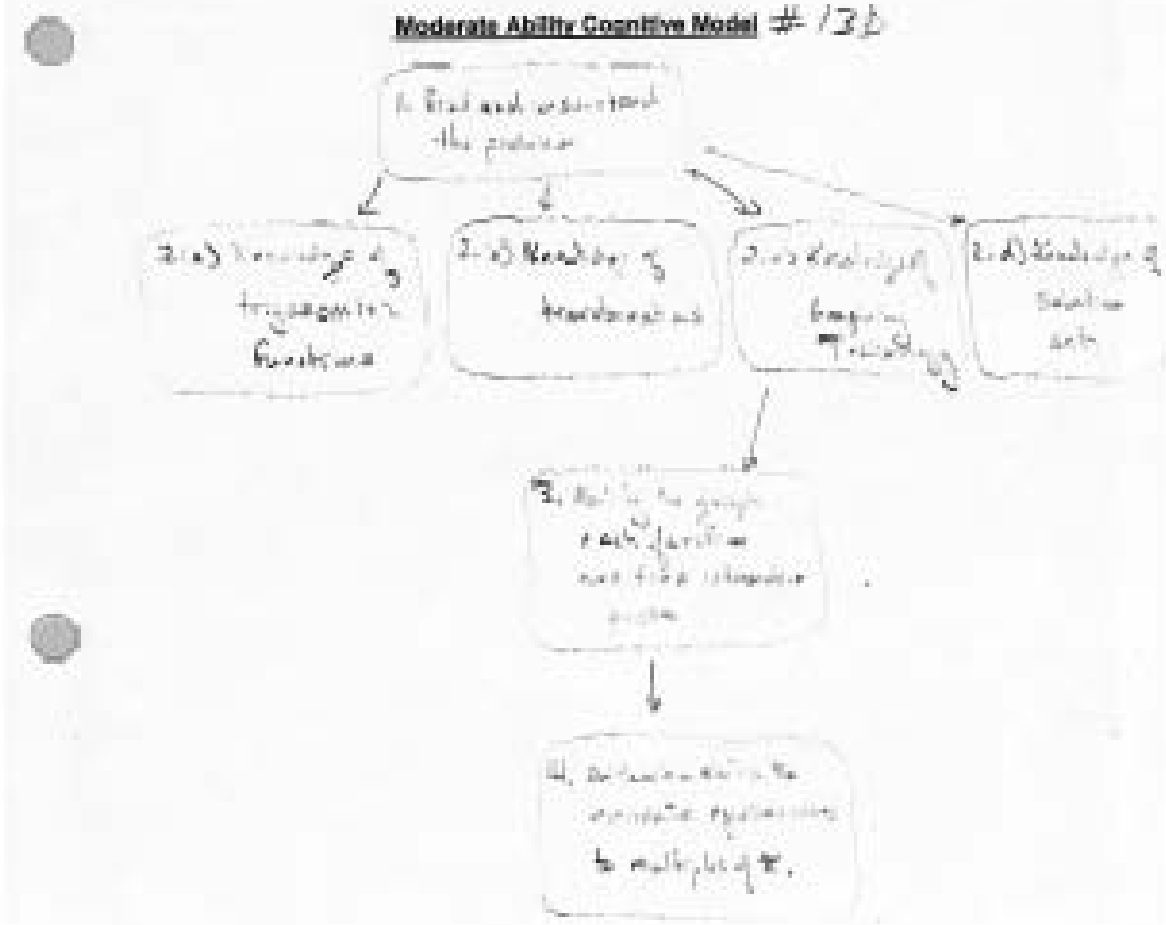

High Ability Cognitive Model #731


Moderate Ability Cognitive Model
1. Read and understand the problem.
2.a) Knowledge of angles in a coordinate plane
2.b) Knowledge of trigonometric ratios
2.c) Graphing skills
2.d) Technology skills
3.a) Trigonometry skills
3.b) Skills with angles on a coordinate plane

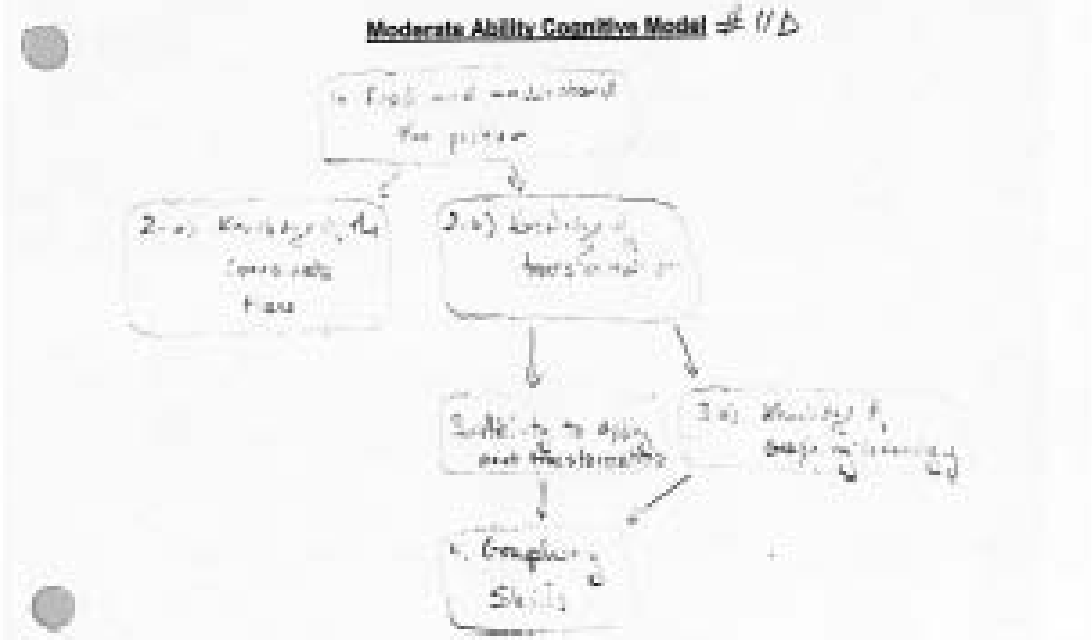

High Ability Cognitive Model
1. Read and understand the problem
2.a) Graphing
2.b) Knowledge of angles in a triangle theorem
3. a) Given the proof, rationale that the angle in the quadrant II and is then 156°


Moderate Ability Cognitive Model # 3
1. Read and understand the problem.
2.a) Knowledge of graphing terms
2.b) Knowledge of translation
2.c) Knowledge of technology
2.d) Algebra skills
3.a) Analyze skills for graph
3.b) Graphing with technology skills
3.c) Solve comparison skills


High Ability Cognitive Model #3
1.
2.
3.
3. b)


Moderate Ability Cognitive Model #13

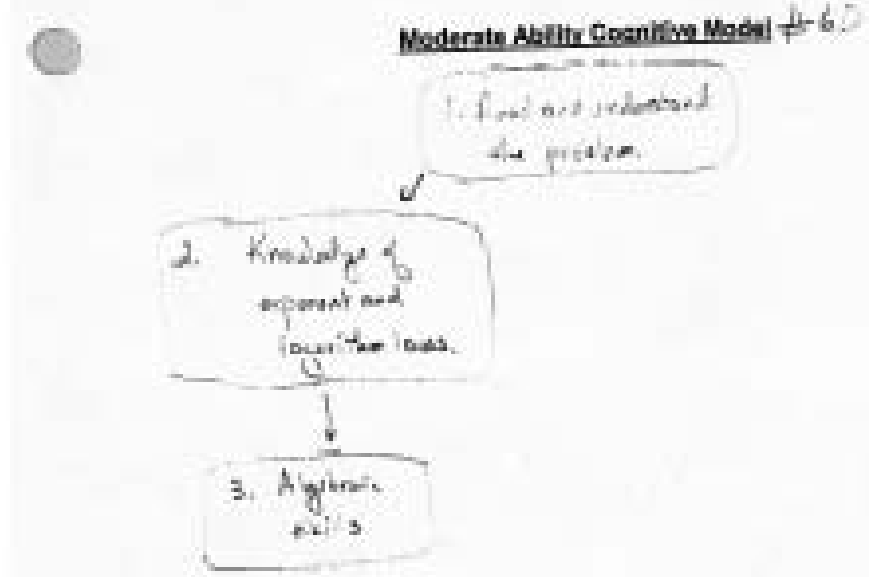
High Ability Cognitive Model #13


Moderate Ability Cognitive Model # 11 b

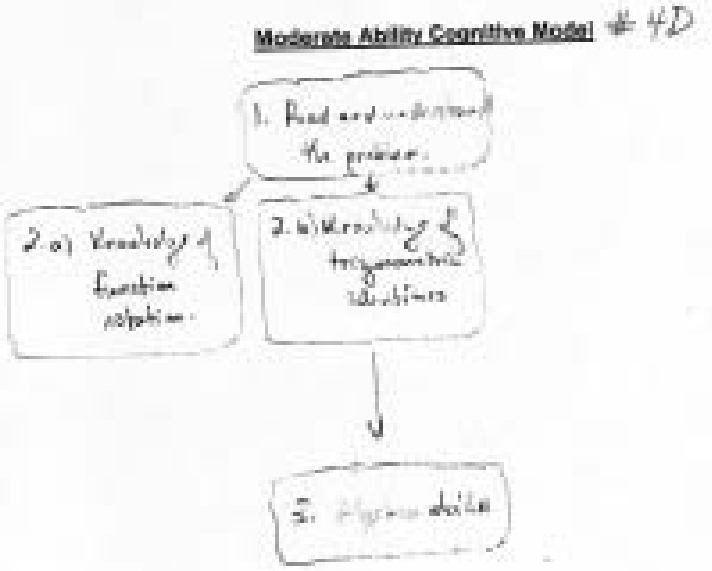

High Ability Cognitive Model #11D


Moderate Ability Cognitive Model #6
1. Find and understand the problem
2. Acknowledge the Fundamental Counting Principle
3. Multiply Fundamental Counting to get # of the options

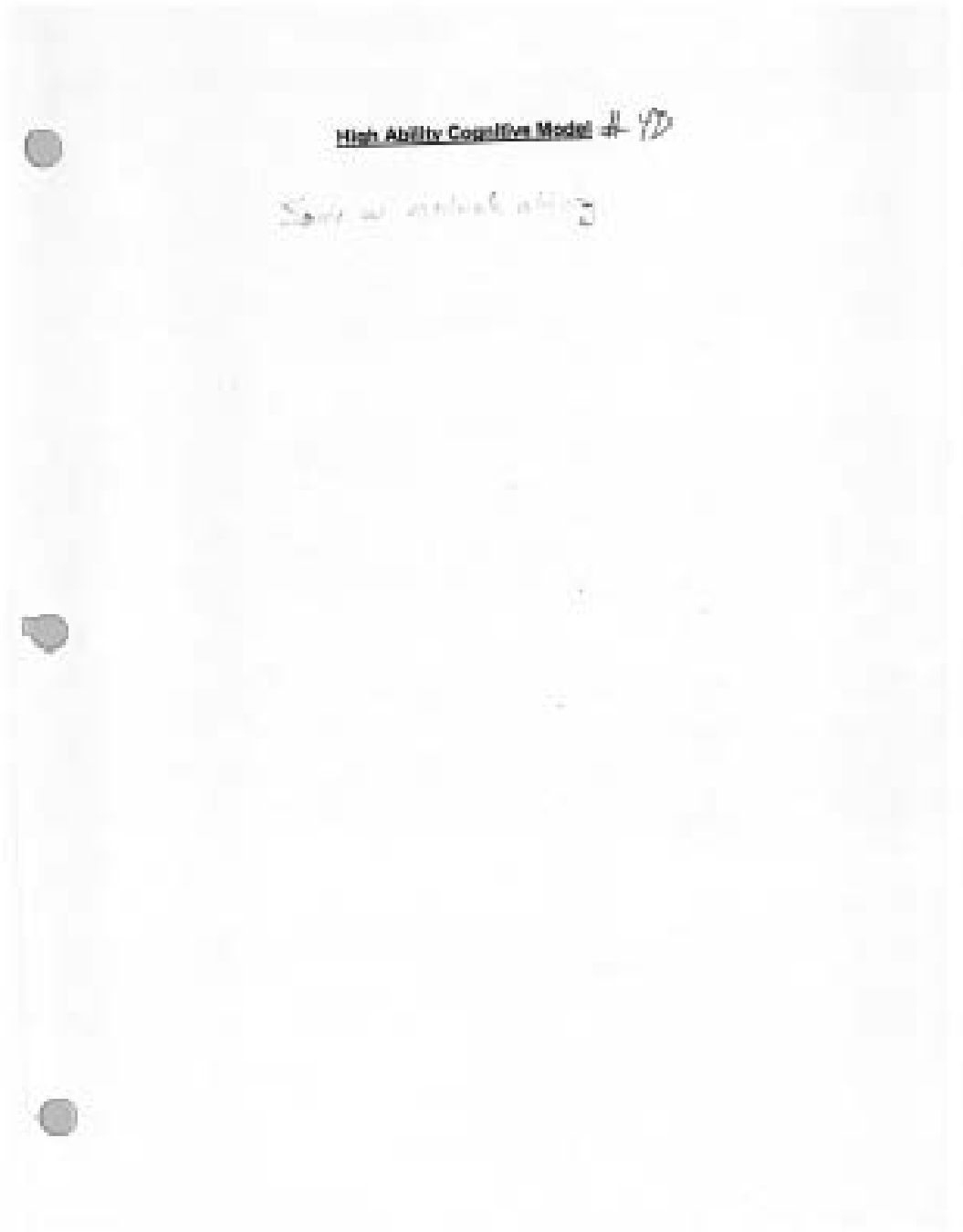

High Ability Cognitive Model #8


Moderate Ability Cognitive Model
1. Find and understand the problem.
2. Knowledge of
3. Algebraic

High Ability Cognitive Model



Note: A student could graph the given function and the alternatives to see which one is correct, but it would still require algebraic skills to match a function to its graph. The student would need knowledge of trigonometric FUNCTIONS (as well) and knowledge of trigonometric IDENTITIES to choose which alternative is correct.


High Ability Cognitive Model #17

Appendix B – Cognitive models (flowchart) of the knowledge and skills that moderate ability and high ability students would use to solve each item correctly, developed by H.R

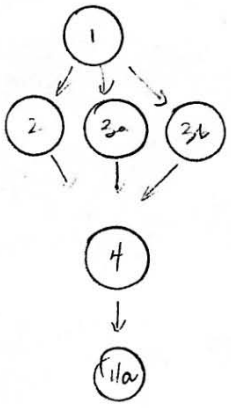
High Ability Cognitive Model #21E

21/5E

High Ability

Would not solve the question differently than a moderate student, although probably visualizes the graph without actually sketching the graph.

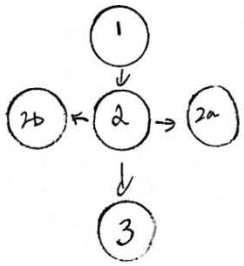
1. Read and understand the question
2. Recognizes that this is an ellipse whose center is (3, -2)
3. Graphing skills
   a. Plots center
   b. Sketches the endpoints of the horizontal axis 2 units on either side of the center
4. Knowledge of domain
   a. Determines that the domain is $1 \leq x \leq 5$



Moderate Ability Cognitive Model #31E

31E
Moderate

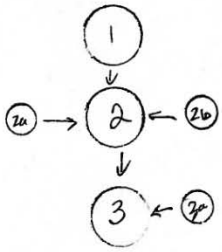
1. Read and understand the question
2. Knowledge of binomial probability
   a. ${}_nC_k p^k(1-p)^{n-k}$
   b. where n=10, k=3, p=0.15
3. Mathematical skills
   a. ${}_{10}C_3 (.15)^3(.85)^7$
   b. 0.129833....
   c. 0.13



High Ability Cognitive Model

31E
High Ability

1. Read and understand the question
2. Knowledge of binomial probability
   a. ${}_nC_k\, p^k(1-p)^{n-k}$
   b. where n=10, k=3, p=0.15
3. Knowledge of calculator skills
   a. Use the calculator feature: bicdf(10, 0.15,3)
   b. 0.13

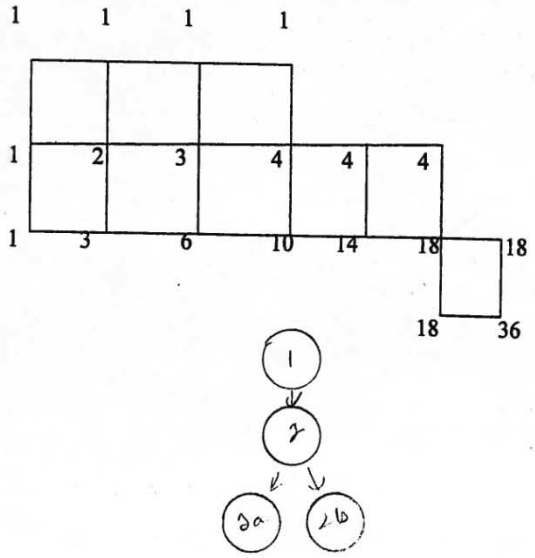


Moderate Ability Cognitive Model #

27E

Moderate

1. Read and understand the question
2. Knowledge that Pascal's Triangle can be used to solve this pathway
   a. Assign to each vertex the numbers in Pascal's Triangle
   b. Check to make sure each diagonal set of values adds to equal value of vertex below and between



High Ability Cognitive Model #27E

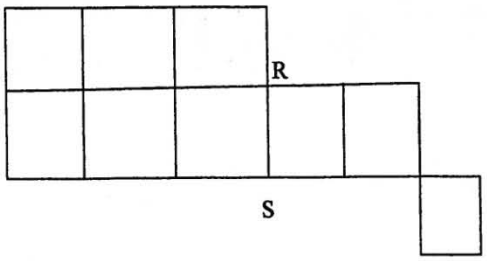
27E

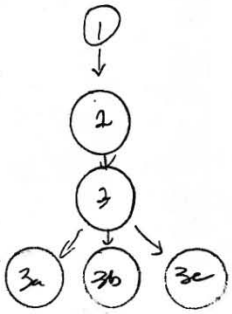
High Ability

A high ability student would use the same model as moderate unless they were not shown that Pascal's Triangle method works for this type of question. An alternate method could be to :

1. Read and understand the question
2. Knowledge to break up the diagram into parts
3. Knowledge of permutations and arrangement of like elements
   a. The number of pathways through R, but not S
      i. $\frac{4!}{3!1!} \times 2 \times 2 = 16$ pathways
   b. The number of pathways through S but not R
      i. $\frac{4!}{2!2!} \times 1 \times 2 = 12$ pathways
   c. The number of pathways through S and R
      i. $\frac{4!}{3!1!} \times 1 \times 2 = 8$ pathways

A total of 36 pathways





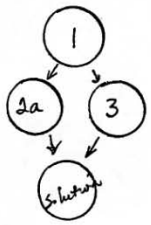
Moderate Ability Cognitive Model 5E

5E

Moderate

1. Read and understand the question
   a. Simplifies the question by eliminating choices
2. Knowledge of transformations
   a. Recognizes that 3 units left limits the choices to B or C
3. Graphing skills that a horizontal stretch of 2 implies f($\frac{1}{2}x$)
   a. Therefore the selection is B

Solution "B"

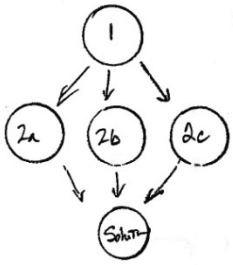


High Ability Cognitive Model # 5E

5E

High Ability

1. Read and understand the question
2. Knowledge of transformations
   a. A reflection in x axis, implies $y = -f(x)$
      i. Every selection is reflected in x axis
   b. Horizontally stretched by a factor of 2, implies $y = f(\frac{1}{2}x)$
      i. Selection A or B
   c. Translated 3 units left, implies $y = f(x+3)$
   d. Checks that choice B is translation down by 1 unit

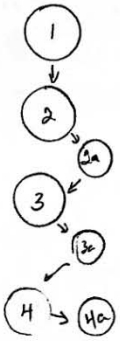


Moderate Ability Cognitive Model

4 E

Moderate

1. Read and understand the question
2. Knowledge of inverse notation
   a. Recognize that x=f(y) is inverse notation
3. Knowledge of ordered pairs
   a. Select key points on the graph such as (-4, -1), (1, -2), (0, -1), (2, -3) and (4, -1)
4. Knowledge of inverse points
   a. Inverse points are (-1, -4), (-2, 1), (-1, 0), (-3, 2) and (-1, 4)
5. Graphing skills
   a. Graph the inverse points by plotting the inverse points on a coordinate plane.

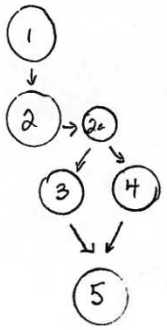


High Ability Cognitive Model #45

4 E
High Ability

This student would probably solve this question the same as a moderate student, however probably not choose as many points on the graph or

1. Read and understand the question
2. Knowledge of inverse notation
    a. Recognize that x=f(y) is inverse notation
3. Knowledge that inverse is a mirror image along the line y=x
4. Knowledge that "horizontal' images become "vertical" in the inverse
    a. So narrow the choices to B or D
    a. Select a key points such as (0, -1)
5. Knowledge of inverse points
    a. Inverse points is (-1, 0), therefore answer is "D"



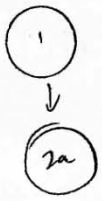
Moderate Ability Cognitive Model #29M

29M

Moderate

1. Read and understand the question
2. Knowledge of Pathways
    a. Physically count the number of different pathways

Solution is C



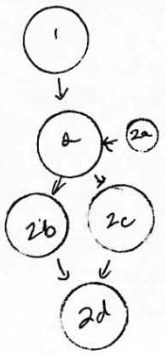
High Ability Cognitive Model #29M

29M

High Ability

1. Read and understand the question
2. Knowledge of Pathways
   a. Recognize that it you have two cubes 1 x 1 x 1
   b. The number of pathways from A to B is $\frac{3!}{1!1!1!} = 6$
   c. The number of pathways from B to C is $\frac{3!}{1!1!1!} = 6$
   d. The number of pathways from A to C is $6 \times 6 = 36$

Solution is C



Moderate Ability Cognitive Model #28M

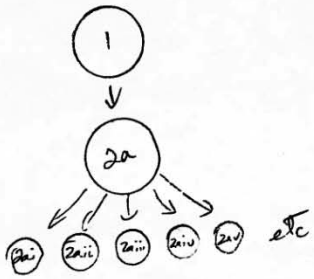
28M

Moderate

1. Read and understand the question
2. Knowledge of combinations
   a. Construct a sample space of outcomes
      i. Team 1 vs 9 other teams = 9 games
      ii. Team 2 vs 8 other teams = 8 games
      iii. Team 3 vs 7 other teams = 7 games
      iv. Team 4 vs 6 other teams = 6 games
      v. Team 5 vs 5 other teams = 5 games
      vi. Team 6 vs 4 other teams = 4 games
      vii. Team 7 vs 3 other teams = 3 games
      viii. Team 8 vs 2 other teams = 2 games
      ix. Team 9 vs 1 other team = 1 games

Total 45 games

Conclude that ${}_{10}C_2$ is equivalent to 45



High Ability Cognitive Model #5 PM

28M

High Ability

1. Read and understand the question
2. Knowledge of combinations
3. Recognize that it takes 2 teams to play a game and that you have 10 teams to choose from.
4. ${}_{10}C_2 = 45$ games

Flow chart
See Key

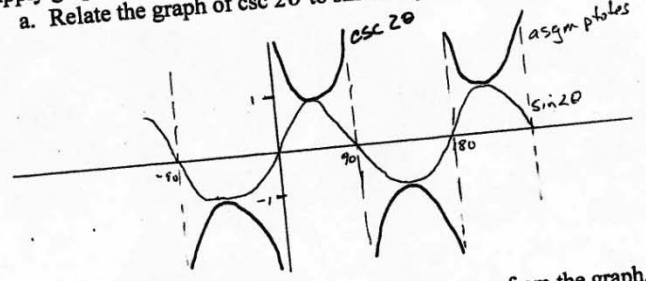
Moderate Ability Cognitive Model # 17M

17M

Moderate

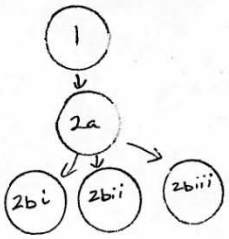
This student would likely use degrees for their sketch rather than radians and then convert into radian measure

1. Read and understand the question
2. Apply graphing skills to knowledge of reciprocal trigonometric functions
   a. Relate the graph of csc 2θ to sin 2θ by using a sketch



   b. Determine the period, range, and asymptotes from the graph.
      i. Period is 180^0 or π
      ii. Range is $y \geq 1$ or $y \leq -1$
      iii. Asymptotes are $0, 90^0, 180^0$ or $0 \pm \frac{n}{2}\pi$

Solution is A



High Ability Cognitive Model # 17M

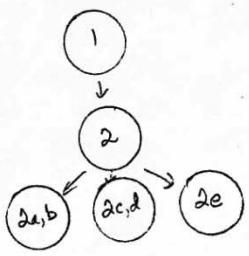
17M

High Ability

This student would likely use their calculator to sketch the graph or simply would be able to visualize what the graph looks like.

1. Read and understand the question
2. Knowledge of reciprocal trigonometric functions and transformations
   a. Recognize the period of csc $\theta = 2\pi$
   b. Recognize the period is a compression by a factor of 2 so
      csc $2\theta = \pi$
   c. Recognize the asymptotes of csc $\theta = 0 \pm n\pi$
   d. Recognize the asymptotes are a compression by a factor of 2 so
      $\csc 2\theta = \frac{1}{2}(0 \pm n\pi) = 0 \pm \frac{n}{2}\pi$
   e. Recognize that the range of csc 2θ is $y \geq 1$ or $y \leq -1$

Solution is A

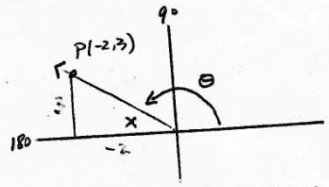


Moderate Ability Cognitive Model #16M

16M

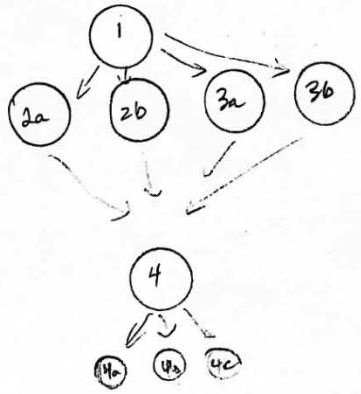
Moderate

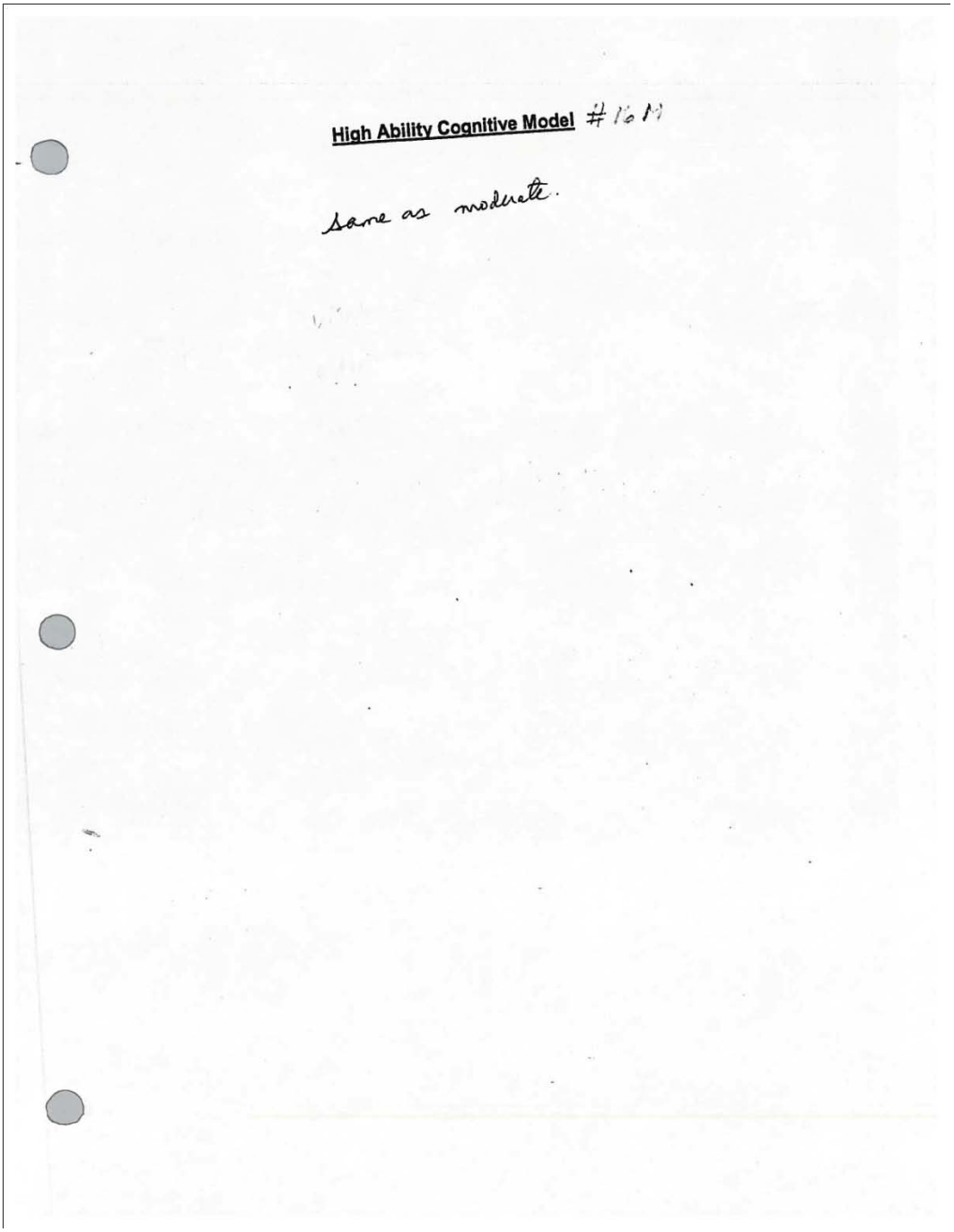
1. Read and understand the question
2. Graphing skills
   a. Draw an angle in standard position, complete the triangle
   b. Label the sides



3. Knowledge of trigonometric functions
   a. Solve for x using $\tan^{-1}\left(-\frac{3}{2}\right)$
   b. $X = 56.3^0$
4. Knowledge of the CAST rule and supplementary angles
   a. Tan function is negative in second quadrant
   b. Recognize that θ and x are supplementary angles
   c. $\theta = 180 - 56.3 = 123.7^0$

Solution is C




High Ability Cognitive Model #16M
Same as moderate.

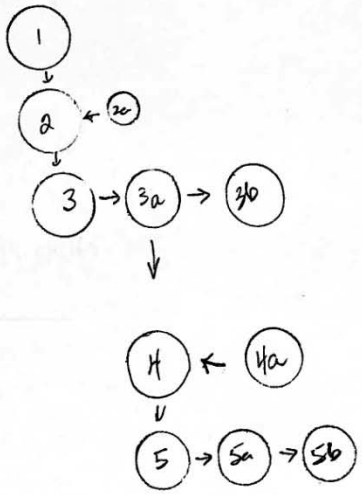
Moderate Ability Cognitive Model #3M

3M

Moderate

1. Read and understand the question
2. Knowledge of conic functions
   a. Recognize the function is a parabola with a vertex (0, 3)
3. Graphing skills
   a. Sketch the parabola
   b. Knowledge of range and domain
      i. the domain is $x \in R$ and $y \geq 3$.
4. Knowledge of transformations
   a. Recognize that the parabola is translated 4 units right and 2 units up
5. Graphing skills
   a. Sketch the translated parabola
   b. Knowledge of range and domain
      i. the domain is $x \in R$ and $y \geq 5$.

Solution is C



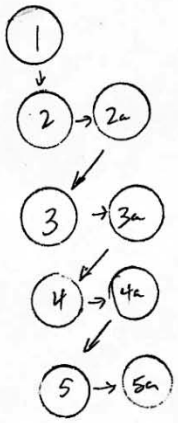
High Ability Cognitive Model #3M

3M

High Ability

1. Read and understand the question
2. Knowledge of conic functions
   a. Recognize the function is a parabola with a vertex (0, 3)
3. Knowledge of domain and range
   a. Visualizes that the domain and range are $x \in R$ and $y \geq 3$.
4. Knowledge of transformations
   a. Recognize that the parabola is translated 4 units right and 2 units up
5. Knowledge of domain and range
   a. Visualizes that the domain and range are $x - 4 \in R$ and $y - 2 \geq 3$ and therefore $x \in R$ and $y \geq 5$.

Solution is C



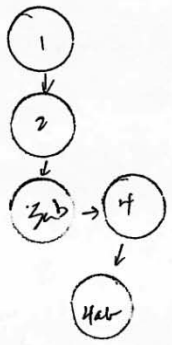
Moderate Ability Cognitive Model # 13D 25

13D

Moderate

1. Read and understand the question
2. Knowledge that the Trig equation can be solved as a system
3. Knowledge of graphing using a graphing calculator
   a. Y = 1
   b. Y = tan (3x)
4. Knowledge that the intersection points of the two graphs are the solutions
   a. Using degrees, the intersection points are 15^0, 75^0, 135^0, etc
   b. Converting to radians: $x = \frac{\pi}{12} + \frac{n\pi}{3}, n \in I$

Solution is D



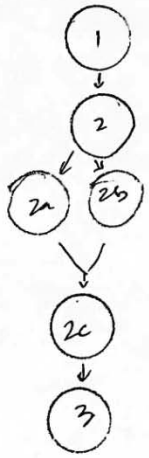
High Ability Cognitive Model # 13D

13D

High Ability

1. Read and understand the question
2. Knowledge of the unit circle
   a. Tan (3x) = 1
   b. $3x = \frac{\pi}{4}, \frac{5\pi}{4}, \frac{9\pi}{4}, \frac{13\pi}{4} \ldots$
   c. $x = \frac{\pi}{12}, \frac{5\pi}{12}, \frac{9\pi}{12}, \frac{13\pi}{12} \ldots$
3. Knowledge of general solution
   a. $x = \frac{\pi}{12} + \frac{n\pi}{3}, n \in I$

Solution is D



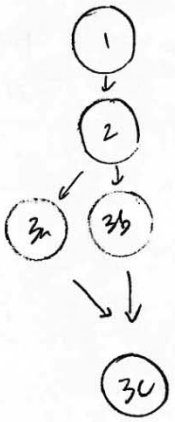
Moderate Ability Cognitive Model #11D

11D

Moderate

1. Read and understand the question
2. Knowledge of transformation
3. Knowledge of vertically stretches
   a. Point G (0,3) is 2 units away from the line y = 1, apply a vertical stretch by a factor of $\frac{1}{2}$ to G gives you G' that is 1 unit form y = 1 thus G'(0,2)
   b. Only choice B has a point G'(0,2)
   c. Check if new vertex is V'(2,0)
      i. Point V (2, -1) is 2 units away from the line y = 1, apply a vertical stretch by a factor of $\frac{1}{2}$ to V gives you V' that is 1 unit form y = 1 thus V'(2,0)

Solution is B



High Ability Cognitive Model # 11D

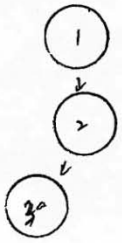
11D

High Ability

1. Read and understand the question
2. Knowledge of transformation
3. Knowledge of vertically stretches
   a. Point G (0,3) is 2 units away from the line y = 1, apply a vertical stretch by a factor of $\frac{1}{2}$ to G gives you G' that is 1 unit form y = 1 thus G'(0,2)

Only choice B has a point G'(0,2)

This type of student would probably not check using the extra point.
Solution is B



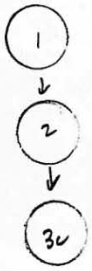
Moderate Ability Cognitive Model #8D

8D

Moderate

1. Read and understand the question
2. Knowledge of permutations
3. Using the fundamental counting principle
   a. The first position on bottom row can be filled with 10 different students
   b. The second position on the bottom row can be filled with 9 different students etc
   c. 10 x 9 x 8 x 7 x 6 x 5 x 4 x 3 x 2 x 1

Solution is D



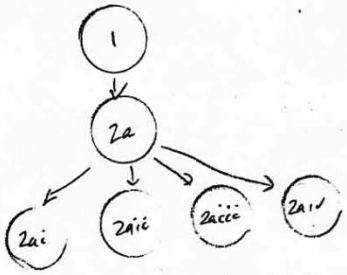
High Ability Cognitive Model #8D

8D

High Ability

1. Read and understand the question
2. Knowledge of permutations
   a. Break into a simpler problem
      i. Bottom row $_{10}P_4$
      ii. 3rd Row $_6P_3$
      iii. 2nd Row $_3P_2$
      iv. Top Row $_1P_1$

Solution is D



Moderate Ability Cognitive Model #6D

6D

Moderate

1. Read and understand the question
2. Replace a & b with arbitrary numbers eg a =10, b = 2
3. Evaluate $\log_{10}(10^4 \times 2) - \log_{10}(10 \times 2) = 3$
   a. Check answers C & D by substituting a = 10.

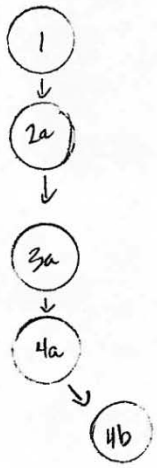
Answer is A

High Ability Cognitive Model #6D

6D

High Ability

1. Read and understand the question
2. Knowledge of laws of logarithms
   a. Combine into a single log
      i. $\log_a\left(\frac{a^4b}{ab}\right)=x$
3. Knowledge of exponents
   a. Simplify exponents
      i. $\log_a\left(a^3\right)=x$
4. Knowledge of how to change from log form to exponential form
   a. $a^x=a^3$
   b. $x=3$

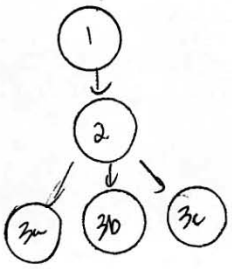


Moderate Ability Cognitive Model #4D

4D
Moderate
A moderate student would have difficulty getting this question correct. This question is not well constructed because the choices are a 3-1 split, so A is probably eliminated as a possible answer.

1. Read and understand the question
2. Knowledge that identities can be proven using verification
   a. Let $\theta = 15^0$
3. Mathematical skill of substitution
   a. The value of $1852.21 + 3.11(\cos(2\theta))$ is 1854.90… (Stem)
   b. The value of choice A: $(1855.32 - 6.22\cos^2\theta)$ is 1849.5…
   c. The value of choice B: $(1849.10 + 6.22\cos^2\theta)$ is 1854.90…

Therefore the answer is B



High Ability Cognitive Model #4D

4D

High Ability

1. Read and understand the question
2. Knowledge of Double Angles Trigonometric Identities
   a. $\text{Cos } 2\theta = \cos^2\theta - \sin^2\theta$
3. Mathematical skills
   a. $\text{Cos } 2\theta = \cos^2\theta - (1 - \cos^2\theta)$
   b. $\text{Cos } 2\theta = 2\cos^2\theta - 1$
   c. Substitute: $1852.21 + 3.11(2\cos^2\theta - 1)$
   d. Expand: $1852.21 + 6.22\cos^2\theta - 3.11$
   e. Collect like terms: $1849.10 + 6.22\cos^2\theta$

Solution is B.