The Effects of Cognitive Ability, Motivation and Perception of Interview on the Accuracy and Validity of Think Aloud Interview Data

by

Alina Lutsyk

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Education

in
Measurement, Evaluation and Cognition

Department of Educational Psychology
University of Alberta

Abstract

The objective of the present study was to examine the impact of examinee variables such as cognitive ability, motivation, and perception of the interview on the accuracy and validity evidence of the think aloud interview, a method widely used to collect information on participants' response processes in assessment validation studies. Think-aloud interviews, which included a series of tasks and questionnaires, were conducted with a sample of 66 undergraduate students in the large research-intensive university. The following research questions were investigated: 1. Do participants' cognitive ability and motivation levels influence the accuracy of problem solving during the think aloud interview? 2. Do levels of anxiety and comfort influence students' perception of the think-aloud interview? 3. Do students with more skeptical perceptions of the think aloud interview take less time to solve tasks than students with less skeptical perceptions; and 4. Do students' anxiety or discomfort as well as attitudes about mistakes influence the accuracy of problem-solving during think-aloud interviews? The results of the present study indicate that cognitive ability and motivation levels did not significantly influence the accuracy of task solutions. Level of anxiety did not have impact on the accuracy either. However, participants' direct level of comfort and behavioral attitudes towards mistakes were significantly related to the accuracy of task solutions. Participants' indirect perception of the effectiveness of the interview was also related to completion time and their indirect level of comfort with the think-aloud interview. Finally, participants' direct perception of the effectiveness of the interview was also related to participants' external motivation and their direct level of comfort with the think-aloud interview. Implications of the results are discussed for best practices in validation studies.

Preface

This thesis is an original work by Alina Lutsyk. The research project, of which this thesis is a part, received research ethics approval from the University of Alberta Research Ethics Board, Project Name "The Effects of Cognitive Ability, Motivation and Perception of Interview on the Validity of Think Aloud Interview Data," No. Pro00073567, July 10, 2017.

The research conducted for this thesis forms part of an international research program being led by Professor J.P. Leighton at the University of Alberta with support from the Social Sciences and Humanities Research Council of Canada. Some of the ideas generated for investigation originated with me (e.g., rationale for inclusion of IQ in relation to response processes) and others with Dr. Leighton (e.g., conceptualization, development of Anchoring Vignettes and ATMI). The literature review, method, data analysis and discussion for this thesis was largely conducted by me with guidance from Dr. Leighton given previous research she has conducted in the domain of think alouds/protocol analysis and cognitive labs. An international peer-reviewed conference presentation was delivered at the International Test Commission meeting in Montreal 2018. The citation is as follows: Lutsyk, A. & Leighton, J.P. (2018, July 5). *The effects of cognitive ability, motivation and perception of interview on the validity of think aloud interview data.* Presentation delivered at the annual meeting of the International Test Commission (ITC) conference. Montreal, Quebec, CANADA. I was responsible for the data collection and part of the analysis as well as generating the first draft of the presentation. Dr. Leighton assisted with the introduction and discussion.

Dedication

This thesis is dedicated to my loving and supportive parents Petro and Natalia Lutsyk who always challenge and encourage me in my educational pursuit.

Acknowledgments

I would like to express my gratitude to my supervisor Dr. Jacqueline P. Leighton for the continuous support of my MEd study and research, for her expertise and insights that widen my knowledge and understanding, for the guidance and mentorship throughout the whole process.

I would also like to acknowledge the committee members Dr. Maria Cutumisu and Dr. Rebecca Gokiert for their valuable feedback and inputs.

Finally, I want to thank my husband Kyle R. King who had to sacrifice his comfortable living in the USA and move to Canada so that I could continue my studies. His unconditional support and commitment allowed me to stay focused on my research and study goals.

Table of Contents

List of Tables

Introduction

According to the Standards for Educational and Psychological Testing (AERA, APA, NCME, 2014), claims made about students from their test results must meet specific evidentiary criteria for quality and fairness. Educational tests such as standardized achievement tests are often used to evaluate and make inferences about students' knowledge and skills. These evaluations and inferences can be consequential for students in influencing their educational and professional opportunities. Often, the aims of achievement tests are not only to evaluate whether students have acquired specific knowledge and skills but also to evaluate whether they have acquired the underlying cognitive response processes that makes their acquired skills operate successfully not just for the test but in the real world. A question that is often asked by stakeholders, psychometricians, educators and policymakers, is whether claims made about examinees from their test-based performance are accurate in what they reveal about what examinees know and can do? Consequently, psychometricians and testing specialists have, specifically, become concerned with the methods used to validate the inferences made from these test scores to ensure that the tests are working as expected in their measure of students' knowledge and skills. The concern is understandable as many standardized achievement tests such as college readiness exams developed by ACT and the College Board in the US, as well as Alberta's Diploma Examinations, are high-stakes and used to make decisions about whether students enter into colleges or university programs of their choice.

Kane (2013) defines the process of validating test-based inferences as an evaluation of the plausibility of the claims based on test scores. In other words, this means that a *validity argument* must be created to show evidence for the specific claims made from test scores. According to the Standards (AERA, APA & NCME, 2014), five sources of evidence for the validity argument must be demonstrated, namely, evidence based on (1) test content, (2) internal

structure, (3) external structure or relations to other variables, (4) consequences of test performance and (5) response processes. Validity evidence based on test content includes the relationship between the subject-matter content (themes, wording, and the format of the test items) and the construct the test is intended to measure. Validity evidence based on internal structure involves an analysis of test items, their interrelationships and homogeneity, and their relationship with the construct on which the proposed interpretation of the scores is based. Validity evidence based on external structure involves an analysis of the relationship of test items with other tests that reflect similar and distinct constructs. This is done to show what a test measures and what it does not measure. Validity evidence based on consequences of test performance is designed to show that the interpretation of test scores is leading to intended uses and outcomes. Finally, validity evidence based on response processes is designed to provide evidence of the unobservable thinking and problem-solving processes test-takers used to answer test items, which are expected of the construct measured by the test items. A sound validity argument integrates these various types of validity evidence to illustrate "the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses" (AERA, APA & NCME, 2014, p. 21).

Objective

The purpose of the research reported in the present thesis is to critically examine evidence based on students' response processes gathered using the think-aloud interview. In particular, the focus is on the *think aloud interview* as this is a method that is often used to collect data on students' responses processes (Leighton, 2017). Although the think-aloud interview originated as a psychological method employed in the development and verification of information-processing theories (Ericsson & Simon, 1993), it has been adopted by testing

specialists as way to help evaluate students' or examinees' response processes in relation to their test performance. In the current study the following examinee (or interviewee) variables were investigated for their effect on the accuracy of problem solving during think aloud interviews: (1) cognitive ability, (2) motivation, (3) attitudes towards mistakes, (4) perception of the interview, and (5) demographic and control variables (e.g., gender and past academic achievement).

The balance of the thesis is divided into four sections. The first section provides a description of the literature pertinent to the think-aloud interview and interviewee variables, which may influence performance during the interview. The second section provides information on the methods used to conduct the present study and collect data. The third section presents the results of the statistical analyses of the data. The fourth section concludes with a discussion, providing an account of findings, limitations, and further recommendations relevant to the research topic.

Think-Aloud Interviews

There are various methods used to measure the response processes activated during problem solving. For example, some methods (e.g., eye-tracking) are concentrated on measuring responses based on the intensity of engagement in thought processes. For example, measuring a participant's gaze during a problem-solving task is designed to evaluate the visual information a participant is attending to while generating a response (Gorin, 2006). Other methods such as use of technology-based devices (e.g. wrist watches that measure heart rate) are used to measure emotional responses during problem solving to evaluate the arousal level of participants as they are engaged in solving a problem (see Leighton, Tang & Guo, 2017).

The current study is focused on the method of the think-aloud interview (Ericsson & Simon, 1993) as a tool to elicit participants' response processes. Although developed as a

psychological tool, it gained popularity as an effective method to help to collect response process data in educational testing situations in samples of elementary, secondary and higher education students (e.g., Bruckener, & Pellegrino, 2016; Leighton, 2004). During think-aloud interviews, participants are instructed to verbalize all that they are thinking, concurrently, as they try to solve a task from the moment they see it until they arrive at a final answer. Once they have arrived at a final answer, often the interviewer will ask follow-up or retrospective questions to probe how participants remembered solving the task. The retrospective and concurrent reports are then compared for congruence. Participants, by vocalizing their thoughts concurrently as they are solving the tasks, are prompted to express the response processes they are employing to answer these tasks and that are active in working memory (Ericsson & Simon, 1993; Leighton, 2017).

The response processes revealed during problem-solving activities are of interest to educational testing specialists (e.g., Gierl, 1997; Katz, Bennett, & Berger, 2000; Leighton, Rogers, & Maguire, 1999; Leighton, 2013; Norris, 1990; Wang, & Gierl, 2011) as many have used the think-aloud interview to investigate test item performance. In particular, the think-aloud interview has permitted testing specialists to investigate examinees' goals, strategies, patterns of thought, and even misconceptions by collecting verbal reports that reveal the task information examinees are considering and working through in their task solutions. In particular, using the think aloud interview, testing specialists attempt to evaluate the degree to which standardized test items (e.g., multiple-choice, constructed response and performance-based tasks included in high school exit exams, college/university entrance exams, certification and licensure exams) elicit expected knowledge and skills in participating examinees.

Think aloud interviews may not only be helpful in revealing the response processes examinees activate in working memory as they solve test items, but also in the explanations they

provide for using specific strategies. However, these two sources of information need to be distinguished as they provide evidence for different claims. For example, Ericsson and Simon (1993) distinguish these two types of verbalization as providing insight into different cognitive processes. For example, the *concurrent verbalization* participants provide when they vocalize their thoughts while solving the task and in generating an answer reveals the content they are attending to in working memory as they work through to a solution. These concurrent vocalizations provide the data that can be used to make claims about the problem-solving response processes used to solve a task. In contrast, the *retrospective verbalization* participants provide when they recall how they solved the task provides data from long-term memory about how they believed they solved the task, including specific strategies. The retrospective verbalizations do not represent the immediate contents of working memory. Rather, these verbalizations represent beliefs about problem solving, which are being retrieved from long-term memory. These beliefs are more likely to contain constructions and other potentially biased elements not reflecting an accurate recollection of the problem-solving approach (Nisbett, & Wilson, 1977; Leighton, 2017).

In comparison to concurrent verbal reports, retrospective verbal reports have been found to be more susceptible to incorrect reporting and bias as students must remember what strategies they used in solving the tasks and describe that recollection (Leighton, 2004, 2013). In a research study comparing both types of verbalizations, Kuusela and Paul (2000) concluded that researchers may want to focus on concurrent verbalization if the objective of the study is to collect as much information as possible about problem-solving. In particular, because the contents of working memory can erode over time, relying on what examinees can remember about their problem solving may not provide the most accurate information needed. However, if

the focus of a study is on explanations or beliefs about the outcome of the task, Kuusela and Paul (2000) suggest that data from retrospective verbalization may yield better information. Nevertheless, it is important to recognize that retrospective data (i.e., verbal reports) may not provide the most defensible support of inferences about examinees' problem-solving given that these data do not reflect the manipulation of information as it is being done in working memory (Leighton, 2017).

Although verbal reports from think-aloud interviews are sometimes used as evidence in validating the interpretations or inferences made from test scores, limited research exists on factors that may impact the accuracy of these data (Norris, 1990; Leighton, 2013, 2017). Aside from the difference between concurrent and retrospective verbalizations, it may also be important to consider other variables that could alter examinee response processes and therefore influence their test item performance. For example, based on their literature review of psychological studies employing think-aloud methods, Ericsson and Simon (1993) argued that specific types of standardized instructions prior to initiating the think aloud interview did not alter or bias the course of response processes; standardized instructions only slowed down the processes. However, Ericsson and Simon (1993) cautioned against probing or querying participants as they were engaged in solving the problem as these interviewer requests for information could potentially bias participants into using, thinking or recalling specific types of information that they would not have otherwise used had they not been probes or asked the questions.

In their review, Ericsson and Simon (1993) also concluded that task characteristics such as task difficulty could impact verbalizations. For example, in one of the studies reviewed, Ericsson and Simon note that verbalization of thinking during easy text comprehension tasks led

participants to read the text out loud without verbalizing any additional information. In contrast, verbalizations of difficult text comprehension tasks (e.g., difficult wording, poor writing) led participants to read slowly and verbalize information that was not in the text.

Prior to Ericsson and Simon's (1993) review of think aloud interviews, Norris (1990) highlighted the need to consider specific procedural factors in collecting data. In particular, Norris (1990) investigated four different ways of eliciting verbal reports from high-school students in comparison to a control group: (1) the think-aloud method involved having participants report all that they were thinking as they worked through the item and selected their final response; (2) the immediate recall method involved having participants explain their selection of a response immediately after selecting a response; (3) the criteria probe method involved having participants indicate immediately after selecting their response whether a specific piece of information in the item made a difference in their selection; and (4) the principle probe method involved having participants indicate immediately after selecting their response whether a general principle in the item made a difference in their selection. The no elicitation method reflected a control group where participants worked through items but did not provide verbal reports.

From his experimental design, Norris (1990) concluded that procedural method of eliciting reports did not alter the response processes reported by participants nor did it affect accuracy. However, the tasks Norris (1990) presented to participants were sampled from a test of critical thinking. The tasks did not include high-stakes achievement test items, which may be more susceptible to eliciting anxiety and efforts to present a socially desirable self. Based on his study, Norris (1990) only cautions that verbal reports must contain sufficient information to compare the content of the reports with participants' chosen answers. Other researchers have also

investigated whether the think-aloud interview can yield data on effects of task format and/ or other factors that might impact accuracy of performance (Katz, Bennet, & Berger, 2000) and whether providing time limits on the length of the interview can impact accuracy (Fox, Ericsson, & Best, 2010; Gilhooly et al., 2010). Katz et al. (2000) showed that task format affected the difficulty of some tasks but not the strategy used by students to solve the tasks; the reverse effect was found with other items. For example, Katz et al. (2000) observed that task format affected students' problem-solving strategies (e.g., comprehension difficulties led to variations in the strategies chosen) but not the overall difficulty of the task for students. Thus, comprehension issues as well as other factors (e.g., students' own inclinations) determined students' strategy choice.

When investigating the effect of timing on the accuracy of verbal reports, Gilhooly et al. (2010) found that imposing time limits prematurely cut off participants' problem solving before they could generate solutions; thus, accuracy of solutions was potentially reduced. Fox, Ericsson, and Best (2010) confirmed these findings, indicating that imposing time limits on think-aloud interviews prematurely cut off participants' problem solving before they were ready to reach their solutions. The study by Fox et al. also confirmed that when participants are given sufficient time in generating task solutions during concurrent verbalizations, the process of thinking aloud did not influence the accuracy of performance. In another study, Leighton (2013) investigated whether interviewer knowledge and item difficulty influenced students' level of nervousness, performance accuracy, and response processes. She found that interviewers' level of expressed expertise in the domain of the task depressed students' accuracy, altered the response processes reported by students during retrospective verbalizations but did not have a measurable impact on nervousness during the interview or as measured by a test of anxiety after the interview.

Interestingly, the tasks Leighton (2013) administered during the think aloud interviews comprised high-stakes achievement math items and differences in response processes were found in contrast to those of Norris (1990).

Interviewee Variables

Much of the previous psychological literature has focused on the procedural factors that could affect the validity of claims made from verbal report data. However, no research has been conducted to evaluate whether human factors or interviewee variables might also influence response processes during the think-aloud interview; and therefore, introduce bias into the content of participants' verbal reports. For example, participants' cognitive ability, motivation, attitudes towards mistakes and perception of the interview could influence the content of working memory and subsequent verbal reports. Consequently, the following interviewee variables were investigated.

Cognitive Abilities

Cognitive abilities can be described in terms of mental factors and skills that may influence performance on tasks. The Penguin Dictionary of Psychology (Reber, Allen, & Reber, 2009) defines cognition as "mental 'behaviours' with underlying characteristics of an abstract nature involving symbolizing, insight, expectancy, complex rule use, imagery, belief, intentionality, problem-solving and so forth" (p.139). As can be noted with this definition, general cognition is often used to denote intelligence and defined broadly, including not only mental activities but also individual thoughts, ideas and knowledge. In fact, even early theorists such as Wundt, James, Spearman, and Thorndike could not agree on the specific cognitive abilities often used to define intelligence. It took the emergence of cognitive psychology to

define cognitive abilities in terms of logical reasoning and abstract thinking (Goldstein, Princiotta, & Naglieri, 2015).

Two of the most influential perspectives on cognitive abilities are based on developmental progressions and/ or psychometric approaches. For example, Piaget (1972) and Ceci (1993) focused on developmental explanations of cognitive abilities (Goldstein, Princiotta, & Naglieri, 2015; Whiston, 2009). Piaget suggested that children move through a series of developmental stages of cognitive abilities by using such functions as assimilation and accommodation. He argued that through assimilation a child incorporates a stimulus into an existing cognitive schema or alters an existing schema through accommodation of the new information. Ceci proposed that context can influence the development of cognitive abilities, and knowledge and skill acquisition in complex environmental domains (Whiston, 2009).

Other theorists such as Guilford, Cattel, Horn, and Carrol provided a psychometric perspective in explaining cognitive abilities. For example, Guilford (Guilford, 1988; Whiston, 2009) proposed a theory that included factors ordered around three dimensions: (1) mental operations (e.g., memory retention, evaluation), (2) content areas (e.g., visual, semantic, behavioural), and (3) products (e.g., units, transformations, implications). The dimensions were proposed as being interactive or compensatory and leading to an ability quotient specific for each person. Another, influential psychometric view on cognitive abilities emerged from the work of three individuals leading to the Cattel-Horn-Carroll model (Whiston, 2009; Evans, Floyd, McGrew, & Leforgee, 2002). Cattel (1971) suggested that cognitive abilities consisted of two factors: (1) fluid abilities were biologically determined and were reflected in memory span skills and spatial thinking; and (2) crystalized abilities were acquired knowledge and skills by means of culture, social interactions, and educational experiences. Horn (1985) later added several other

intelligences, which Carrol (1997) assembled into a hierarchical structure. Overall, efforts to identify the main components of cognitive abilities have led to the recognition that speed of information processing and working memory are core components (Goldstein, Princiotta, & Naglieri, 2015).

Various tests (i.e., Wechsler Scales, Stanford-Binet Intelligence Scale, Slosson Intelligence Test) have been used to measure cognitive abilities. However, it is important to note that educators distinguish between tests of cognitive abilities and achievement (Winston, 2009). On the one hand, tests of cognitive ability are designed to measure general capacities such as thinking abstractly, solving problems, understanding complex ideas, and learning new material. On the other hand, achievement tests are designed to measure acquired knowledge or proficiency in a subject content domain such as Language Arts or Math. Achievement tests measure the extent to which an individual has acquired a certain amount of information or mastered certain skills.

Some research has shown that there is a significant relationship between cognitive abilities and students' academic achievement as measured by GPA or a grade for a single course (Goldstein, Princiotta, & Naglieri, 2015; Farsides, & Woodfield, 2003; Murray, & Wren 2003; Ridgell, & Lounsbury, 2004). Naglieri in his Chapter on Assessment of Intelligence (Goldstein, Princiotta, & Naglieri, 2015, pp.296-316) compared cognitive ability and achievement tests and found that questions on both types of tests were very similar. Notwithstanding their similarity, the constructs of cognitive ability and academic achievement are distinct as they do not share a perfect correlation. Cognitive ability was a variable of interest in the current study as it could influence the accuracy of problem solving or even the motivation experienced by students during the interview. For example, students of higher ability might find academic tasks more interesting

and expend more effort on solutions than students of lower ability. No research to date has been conducted to see if students' performance on achievement items used in think-aloud interviews can be associated to their cognitive abilities. For this reason, the Draw a Person Intellectual Ability Test (DAP: IQ) was used in the present study as it is supposed to be minimally affected by culture and is developmentally relevant (i.e., cognitive ability scores are adjusted for participant age). Moreover, the DAP:IQ has been shown to be significantly correlated with other IQ and achievement tests (Reynolds & Hickman, 2004).

Motivation

Behavioural, psychodynamic, and cognitive approaches have been used to explain motivation throughout history (Ryan, 2012). Motivation in psychology has been defined as an "intervening process or an internal state of an organism that impels or drives it to action" (Reber, Allen, & Reber, 2009, p. 487). Contemporary cognitive theories of motivation suggest that thoughts, beliefs, and emotions are the main processes that underline motivation. For example, Ryan (2012) in his discussion of self-determination theory indicates that people have psychological needs, and fulfilment of these needs facilitates intrinsic or self-directed (autonomous) motivation. Failure to satisfy psychological needs hinders autonomous motivation, leading to more extrinsic motivation (externally controlled) or even lack of motivation (Ryan, 2012).

More specifically, Ryan and Deci (2003; Ryan, 2012) explain both intrinsic and extrinsic motivation. People are believed to exhibit intrinsic motivation when they engage in an activity for the enjoyment and satisfaction of that activity. If people are pressuring themselves, feeling anxious, and working with great urgency, then extrinsic motivation is more likely to be involved

in completion of the task. Extrinsic motivation is also induced by rewards or punishment upon successful completion of the task.

Students' intrinsic and extrinsic motivation has been found to influence their achievement and success (e.g., Covington, 2000; Deci, Koestner, & Ryan, 2001; Lina, McKeachieb, & Kim, 2003; Mauro, 2007). For example, higher levels of students' intrinsic motivation have been found to be positively related to grades, and moderate levels of extrinsic motivation and higher levels of intrinsic motivation has been found to lead to better performance on tasks (Lina, McKeachieb, & Kim, 2003). Wigfield, Cambria, and Eccles (Ryan, 2012, pp.463-478) reviewed a vast number of studies pertaining to students' motivation and achievement. Among some of their conclusions, they indicate that teachers' behaviours toward students, classroom environment, task-level practices, rewards, criteria for success, peers and other factors influence students' motivation and achievement.

Aside from intrinsic and extrinsic motivation generally, researchers are often interested in what specifically moves people to act in certain situations. Guay, Vallerand, and Blanchard (2000) define situational motivation as the volition experienced by students when they are engaged in an activity in the moment. Ryan (2012, p. 463) indicates that it is "the engagement and interest students have in different academic activities, the choices students make about which academic activities to do, their persistence at continuing the activities, and the degree of effort they expend" that needs to be considered in understanding different levels of motivation for specific activities. Guay, et al. (2000) also summarized *amotivational* behaviours as "the least self-determined because there is no sense of purpose and no expectations of reward or possibility of changing the course of events" (p.177). This absence of motivation is evident when people experience a lack of contingency between their engagement in an activity and outcomes. Guay et

al.'s (2000) situational motivation scale was used in the current study to investigate the relationship between motivation specific to solving of tasks during the think-aloud interview.

Perception of the interview

Perception involves the processes that "give coherence and unity to sensory input" and "covers the events from the presentation of physical stimulus to phenomenological experiencing it. Included are physical, physiological, neurological, sensory, cognitive, and affective components" (Reber, Allen, & Reber, 2009, p. 566). Although an increasing number of studies have examined the effect of test anxiety and other emotions (e.g., DordiNejad et.al, 2011; Zeidner, 2007; Pekrun, Goetz, Titz, & Perry, 2002; Ruthig et.al., 2008) on test performance, studies of students' perceptions about testing environments are rare. For example, Ruthig et.al. (2008) investigated the relationship between perceived academic control, emotions (e.g., boredom, anxiety, and enjoyment) and students' achievement, and found that emotions moderated the relationship between perceived academic control and achievement. Other studies have investigated the role of student perception of specific assessment-related factors on test performance. For example, Hong (1999) showed that students' perception of test difficulty had a significant positive effect on their arousal of worry anxiety and emotionality, which influenced test performance. Students who perceived the test to be more difficult expressed higher worry and emotionality in contrast to their peers who perceived the test to be less difficult. The worry anxiety factor, but not emotionality, had a negative effect on students` test performance. Emotionality recalled after test, on the other hand, had a significant positive relationship with performance, although the magnitude was small. However, the pattern of relationship between emotionality and performance has not been consistent, and it has varied from no relationship to a weak negative relationship, and to a positive relationship. Morse and Morse (2002) investigated

students' perception of item difficulty on their performance and found that students performed better on those items they perceived to be less difficult. In another study, Church, Elliot, and Gable (2001) found that students' perception of classroom environment (e.g., lecture engagement, evaluation focus, and harsh evaluation) impacted their adoption of achievement goals, which, in turn, influenced their performance and intrinsic motivation. For example, perceived lecture engagement was found to be a positive predictor of mastery goals whereas perceived evaluation focus and harsh evaluation were negative predictors. Perceived lecture engagement displayed a positive relationship with intrinsic motivation and performance whereas perceived evaluation focus and harsh evaluation displayed a negative relationship. In addition, a perceived evaluation focus was found to be a positive predictor of both performance-approach and performance avoidance goals. However, only perceived harsh evaluation was found to be a positive predictor of performance avoidance goals.

There may be factors that influence students' perception of the think-aloud interview, and these perceptions may influence their performance especially when the interview is focused on academic test items. For example, students might perceive the think-aloud interview as a testing situation and become nervous or anxious, thus possibly interfering with their thinking and performance. In the current study students were asked to complete a series of anchoring vignettes (Kyllonen & Bertling, 2013) designed to measure their perceptions of the think-aloud interview as an effective vehicle for measuring their cognitive response processes and strategies in solving math-type achievement tasks. No studies have been conducted to investigate students' cognitive and emotional perceptions of the think-aloud interview.

In relation to students' perception of the interview, it was also considered worthwhile to evaluate students' attitudes towards their academic mistakes (Leighton, Tang, & Guo, 2018).

Students' attitudes towards their mistakes is an under-researched topic. Yet, most people do not like making mistakes publicly and this could influence students' willingness to report their response processes as they solve academic tasks during the think aloud interview. In the current study, the Attitudes Towards Mistakes Inventory (ATMI; Leighton et al., 2018) was used to measure students' attitudes towards mistakes to determine if these attitudes were related to participants' accuracy of problem solving during the think aloud interview. In the next section, the specific research questions guiding the study as well as the methods used are described.

Method

A non-experimental correlational design was implemented for the study. Aside from the order of math-type tasks presented to participants, no variables were manipulated. The following four research questions were examined in the current study: 1. Do participants' cognitive ability and motivation levels influence the accuracy of problem solving during the think aloud interview? 2. Do levels of reported anxiety and comfort influence participants' perception of the think-aloud interview? 3. Do participants with more skeptical perceptions of the think aloud interview take less time to solve the tasks than students with less skeptical perceptions?; and 4. Do participants' reported anxiety or discomfort as well as attitudes about mistakes influence the accuracy of problem-solving during think-aloud interviews?

Participants

Participants comprised a sample of convenience, which included 66 undergraduate students (28.8% male and 71.2% female) between ages 18 and 45 (M = 23.64, SD = 5.9, Mode=19, Median=22) from a large research-intensive university. Students were part of the Educational Psychology Participant Pool on campus and were enrolled in an Introduction to Educational Technology course. Students were invited to participate in a study of their choice to

gain credit towards their program of study. If students did not wish to participate in a study, they could choose to complete an assignment designed to take an equivalent amount of time. Participating students had not previously been involved in a study involving think-aloud interviews. Fifty six percent of students self-identified as Caucasian/White, 12% as Asian, 6% as Black/ African, 3% as Filipino, 3% as Hispanic, and 1.5% as Aboriginal/Indigenous. Another 3% preferred not to disclose their ethnicity and a remaining 15% identified themselves as Canadian but did not disclose their ethnicity. Forty-five students participated in the study during the Fall 2017 semester and 21 students during the Winter 2018 semester. From here onward, the terms *participants* and *students* are used interchangeably to denote individuals who participated in the present research.

Materials

Think-aloud interviews were conducted with 66 students individually. Students were presented with a booklet that contained three parts. Students were instructed to complete each section of the booklet in the order presented to them. The math-type tasks (booklet part one) were counterbalanced and organized into four different orders labeled as Condition 1 (i.e., games, algebra, logic and statistics), Condition 2 (i.e., algebra, logic, statistics, and games), Condition 3 (logic, statistics, games, and algebra), Condition 4 (statistics, games, algebra, and logic) to avoid a task order effect on accuracy during problem solving. The order of items for the second and third parts of the booklets was the same for all the students. In the next section, each booklet is described.

Booklet part one

Part one contained four multiple-choice, math-type tasks ranging from easy to moderate difficulty. The four math-type tasks were taken from Alberta Education's (2013, 2016)

Mathematics 30-2 Released Items. The tasks were publicly released by Alberta Education from the Diploma Examination Testing Program, which is a summative end-of-year multiple choice high school exit exam. These academic tasks are high-stakes and based on the high school Mathematics 30-2 curriculum, which is designed to provide students with mathematical understandings and critical-thinking skills expected for post-secondary academic programs not requiring calculus. The math-type tasks were sampled to vary in content, and included tasks on games, algebra, logic and statistics, and represented a range of difficulty levels (*p* values ranging from .51 to .98). Each math-type task was presented on a separate sheet (see Appendix 1), allowing students to use the empty space on the page for their notes if needed. In the instructions to the study, students were asked to circle their final answer on the page for each of the tasks. Additional math-type tasks were given to students, when needed, to satisfy the time requirements for credit participation but these were not used in further analyses. As is indicated later in the paper, students were asked to think aloud only during booklet part one of the study.

Booklet part two

Part two included the Draw a Person Intellectual Ability Test for Children, Adolescents, and Adults (DAP: IQ), which is a measure of the way people process information. The test was developed using contemporary psychometric methods. The examiner's manual (Reylolds & Hickman, 2004) for administering the DAP: IQ test indicates that the test has a reliability coefficient (Cronbach's alpha) between $r = 0.80$ and 0.86 (SEM=4) for age groups from 18 to 45 years. In addition, the measure has been validated for use in measuring cognitive abilities. The DAP: IQ involves asking respondents to draw a picture of a full figure of themselves from a frontal view. In doing the DAP: IQ, participating students were provided with paper, pencil, and eraser and were given as much time as they needed to complete the DAP: IQ. Each of the

drawing features of the figure (e.g., eyes, arms, legs) were scored individually according to the DAP:IQ Administration/Scoring Form using 0 to 1, 0 to 2, 0 to 3, or 0 to 4-point scales depending on a specific feature. For example, the eyes are scored in the following manner: 0 = "None present," 1 = "2 eyes present," 2 = "2 oval shaped eyes with pupils," 3 = "meets 2 pts. [and eyes] are proportionately placed on top of face and are grossly symmetrical." Twenty-three features were scored with a maximum of 49 possible points. Students did not think aloud during booklet part two.

Booklet part three

Part three contained four surveys, including two questionnaires, a set of anchoring vignettes, and a final demographic information sheet. The first 16-item questionnaire, called The *Situational Motivation Scale (SIMS)* aimed to measure students' motivation to engage in the think-aloud interview activity (Guay et al., 2000). Shown in Appendix 1, the SIMS comprises four sub-scales: intrinsic motivation, identified regulation, external regulation, and amotivation. Intrinsic motivation concerns behaviors in which people are engaged in for the sake of the activity. Identified regulation occurs when a behavior is perceived as chosen by oneself, yet the activity is not performed for itself but as the means to reach a final goal. External regulation focuses on behaviors a person feels obligated to engage in which are driven by rewards or to avoid negative consequences. Amotivation occurs when the individual feels there is no sense of purpose or no reward associated with the activity performed or its outcome. Students responded to the SIMS statements using a 7-point Likert type scale ranging from 1= "not at all" as the reason for engaging in the activity to 7= "exactly" as the reason for participating in the activity.

Next, a 26-item questionnaire called the *Attitudes Towards Mistakes Inventory (ATMI*; Leighton, Tang & Guo, 2015 [28-item version], 2018 [26-item version]) was administered. As

shown in Appendix 1, the ATMI is intended to measure students' attitudes towards mistakes during learning activities, including discussions and assessments. The questionnaire comprises three sub-scales: affective, behavioral and cognitive. The affective subscale focuses on students' emotional experience; the behavioral subscale focuses on the possible actions students may take when a mistake is made during learning; and the cognitive subscale draws attention to the beliefs students holds about mistakes (Leighton et al., 2018). The 26-item ATMI has subscale internal consistencies of α=.93 for the affective subscale, α = .82 for the behavioral subscale, and r = .82 for the cognitive subscale with adequate external validity (Leighton et al., 2018). Students responded to statements using 5-point Likert type scale ranging from 1 = "strongly disagree" to 5 = "strongly agree."

The next survey administered to participating students included six anchoring vignettes designed to measure their perceived anxiety, comfort and evaluation of the think aloud interview for effectively measuring cognitive processes. As shown in Appendix 1, anchoring vignettes are used to indirectly gauge students' perceptions of an activity as well as their use of rating scales. The benefit of using anchoring vignettes to measure potentially sensitive responses is that they are designed to reduce bias originating from response style and social desirability (Kyllonen & Bertling, 2013). The anchoring vignettes involved having students rate six hypothetical examples of think-aloud interviews before being asked to directly rate their own perception of the interview in which they participated. For example, after each vignette, students responded to three separate items asking about their anxiety and comfort as they thought aloud during their problem solving of the math-type tasks, and finally the effectiveness of using the think aloud interview. For item measuring perceived anxiety, students responded using a 5-point scale ranging from 1=none to 5= extreme; for the item measuring perceived comfort, students again

responded using a 5-point scale 1=very good to 5=very bad; and for the item measuring perceived effectiveness, students responded again using a 5-point scale from 1=completely effective to 5=not effective at all. Each of the six vignettes was presented on a separate sheet concluding with students directly rating their own experience and perception of the think-aloud interview.

Finally, a demographic information sheet was presented to students requesting information about their gender, date of birth, ethnicity, major, year of study at the university, high school GPA, and last year's university GPA.

Procedure and Design

The procedure with the variables of interest used in the study are described as follows.

Procedure

Students signed up to participate in the study through the University of Alberta Participant Pool electronic system where the study was shortly explained to them. Students were asked to not sign up for the study if they have previously participated in the study. The author was trained to administer the interviews and conducted all 66 think-aloud interviews. All participants were asked to sign consent forms. As shown in Appendix 2, the author (functioning as a research assistant) began each interview by self-identifying as the interviewer and reading the instructions from a script to ensure each participant received the same standardized instructions. The instructions contained a short description of the study and explained the objectives and process of the interview to students. Students were advised that they could stop participating at any point if they wished not to proceed. The author asked participants if they understood the instructions and what was expected of them. Students were then informed that they would be audio-recorded during the process of thinking aloud while solving the tasks. Time

of completing the math-type tasks, measured in seconds units, was also recorded. Before participants proceeded to do the actual think-aloud interview they were presented with two short practice tasks on a separate sheet and asked to think aloud to practice the process.

Once the students indicated they understood what was expected of them and indicated comfort with thinking aloud, and finished the practice tasks, they were given a booklet with the math-type tasks and surveys. At this time, the author started to record time from the moment a student started to read or work on the math-type tasks until he or she indicated they were done with all tasks. During participants' thinking aloud, the author did not answer any participant questions related to solving the math-type tasks so as to not influence the cognitive response processes and strategies participants were using to solve the tasks. The participants were asked to let the author know, as they thought aloud, once they had arrived at their final answer. If participants were silent for 10 seconds or longer while thinking-aloud they were asked to "Please, keep talking." If they spoke too softly they were asked to "Please, speak louder." After participants completed each math-type task by identifying an alternative or solution on the page, they were asked to think retrospectively and to provide a verbal report to five questions. The five retrospective questions included the following:

1. Can you tell me the steps or any strategies you used to solve the task?

2. Did you find the task difficult or easy or so-so?

3. Did you feel any emotions as you were solving the task?

4. Is there anything you would change about the task?

5. How difficult or easy was it to solve the task while voicing your thoughts?

All interviews were conducted in a quiet room at the university. Students were requested to think-aloud concurrently only when solving the math-type tasks and not during completion of DAP:IQ, surveys, questionnaires and vignettes. Once participants finished solving the math-type tasks they were instructed to proceed to the DAP: IQ task. Again, the standardized instructions were read from the Administration/ Scoring rubric of the DAP:IQ for each student. After they completed the DAP: IQ they proceeded to fill out the surveys, including the questionnaires and vignettes. Students were invited to ask questions if necessary while working on booklets part two and three (i.e., DAP:IQ, surveys and vignettes). There was no time limit to complete the math-type tasks or the surveys. Interviews lasted 45 minutes to 2 hours.

Overview of variables of interest

The present study focused on human factors that are associated with the accuracy of the responses provided during the think aloud interview. The total number of correct and completed math-type tasks was calculated and used to create an overall accuracy score for each participant. Also for each participant, raw scores from the DAP: IQ test were converted into standardized scores of estimated cognitive abilities (M=100, SD=15). Data on participants' motivation to engage in think aloud during the solving of the math-type tasks and attitudes towards mistakes were summarized by calculating total scores for each sub-scale on the SIMS and ATMI, respectively. Scores for each of the sub-scales were then treated as continuous variables in further analyses. The scores for ATMI items #10, 13, 15, 18, 25, and 26 were reversed to reflect the same direction of the remaining scores within sub-scales. Data on participants' indirect perceptions about anxiety, comfort and effectiveness related to the think-aloud interview while engaging in solving the math-type tasks were summarized. For example, total scores for each of anxiety, comfort, and effectiveness were calculated by adding the scores across all six scenarios

that pertained to a given perception or feeling (e.g., anxiety). Data on participants' direct experiences and perceptions with engaging in think aloud interviews were treated separately and not considered in the calculation of total scores for the anchoring vignettes.

Results

All variables were initially examined for accuracy of data entry, missing values, and measurement scales. There were only two students who did not answer one question on the ATMI questionnaire and only 15.15% of students did not provide their University GPA. Although both qualitative and quantitative data were collected, only quantitative data were analyzed for the present purpose and scope of this study. The following quantitative variables were considered for the present research results: time, accuracy of selections (solutions) of math-type tasks, IQ, university GPA, perceived anxiety (indirect & direct), perceived comfort (indirect & direct), perceived effectiveness (indirect & direct), sub-scores of SIMS (motivation), and subscores of ATMI (attitudes toward mistakes). To reiterate, indirect measures of perceived anxiety, comfort and effectiveness were measured with the six hypothetical scenarios of the anchoring vignettes; whereas direct measures of perceived anxiety, comfort and effectiveness were measured with single items (see Appendix 1). In the following section, preliminary analyses are reported, namely correlations, between variables. Based on these preliminary analyses, variables that were found to be moderately or strongly associated with predicting (a) accuracy on the four math-type tasks and (b) participants' perceived effectiveness (both indirect and direct) of the think-aloud interview were considered for regression analyses.

Preliminary Analyses

Time, accuracy, DAP: IQ, university GPA

Time spent on the four math-type tasks was recorded based on the audio-recording of the think-aloud interview as indicated in the Method section. As shown in Table 1, students spent on average 11 min 2 sec. (M=658.48 sec., SD = 283.61) completing the four math-type tasks with some students completing them quickly in 3 min 27 sec. (207 sec.) and some taking as long as 27 min 11 sec (1631 sec.). In terms of overall accuracy on the four math-type tasks, 28.8% of students provided a correct response to all four tasks, 21.2% answered three tasks correctly, 36.4% correctly responded to two tasks, and 12.1% answered only a single task correctly. Finally, 1.5% of students did not provide a correct answer to any of the tasks. There was no significant correlation between time and accuracy of performance on the math-type tasks (r_s =-.032, n=66, p=.797 >.05).

Table 1
Means and Standard Deviations of Time, Accuracy, IQ, GPA

	Time	Accuracy	IQ	GPA
M	658.48	2.64	103.76	3.3513
SD	283.619	1.076	7.363	0.47520
Min.	207	0	80	2.00
Max.	1631	4	122	4.20

Note. Time metric is in seconds.

Student cognitive ability was also measured with the DAP: IQ. Results indicated that the average IQ was 103.76 (SD=7.36), ranging from 80 to 122. No significant correlation was found between this measure of IQ and accuracy as measured by the number of math-type tasks solved correctly (r_s =.049, n=66, p=.444 >.05). However, a Spearman correlation showed a moderate

but significant correlation between IQ and time (r_s =.308, n=66, p=.012), indicating that participants with higher IQ scores spent more time completing the math-type tasks.

Fifty-six students provided their university GPA scores. The average self-reported university GPA was 3.35 (SD=0.475) using a four-point scale with the lowest reported value being 2.0 and the highest being a 4.2. Originally, students reported their university GPA scores based on either points or percentages depending on the province or country in which they went to school. Because these distinct scales did not provide a common metric, all participant-reported GPA values were placed on a common GPA point scale used in North America (see Appendix 3 for conversion information). There was no significant correlation between students' reported university GPA and accuracy of performance on the math-type tasks (r_s =-.063, n=56, p=.642 >.05).

Anchoring Vignettes: Indirect measures of anxiety, comfort, perceived effectiveness

Six vignettes were administered to students to indirectly evaluate their perceived level of anxiety and comfort with the think-aloud interview (see Table 2 for descriptive statistics), as well as their perceived effectiveness of the think-aloud interview. The scores or ratings ranged from 6 to 30.

Table 2
Means and Standard Deviations of Vignettes Scales

	Vignettes (indirect)			Vignettes (direct)		
	anxiety	comfort	effectiveness	anxiety	comfort	effectiveness
M	16.65	19.53	17.17	2.27	3.89	3.47
SD	3.184	2.494	3.011	0.904	0.914	0.948
Min.	11	13	6	1	2	1
Max.	25	25	26	5	5	5

For anxiety, lower scores indicated that participants did not consider that the characters in the vignettes had experienced much anxiety related to think aloud interviews and higher scores indicated that participants perceived that the characters had experienced anxiety. Participants' reported scores on perceived anxiety ranged between 11 and 25 and the scores were normally distributed. For comfort, lower scores indicated less comfort and higher scores indicated more comfort in the perception of the think aloud interview. Participants' scores ranged from 13 to 25 and were also normally distributed. For effectiveness, lower scores again indicated that participants perceived think-aloud interviews to not be less effective for measuring cognitive processes and higher scores indicated they perceived it more effective. Participants' scores ranged from 6 to 26 and were also normally distributed. A Spearman correlation showed that there was a strong positive correlation (r_s=.574, n=66, p<.000) between participants' indirect perception of comfort and effectiveness with the think aloud interview. Moreover, participants' indirect perception of comfort and effectiveness was negatively correlated with one of the SIMS subscales, namely, the external motivation sub-scale (r_s=-.397, n=66, p=.001 for perceived comfort and r_s=-.370, n=66, p=.002 for perceived effectiveness). In other words, higher levels of perceived comfort and effectiveness of the think-aloud interview were associated with lower levels of external motivation or feeling that external sources are driving the reason for engaging in the activity. A moderate negative correlation (r_s=-.277, n=66, p=.024) was also found between participants' perceived comfort with the think aloud interview and the amotivation sub-scale of the SIMS. Again, this result indicates that students who perceived the think-aloud interview as a comfortable activity did not report lacking in motivation for the activity. Finally, participants' perceived effectiveness of the think aloud interview (r_s=.287, n=66, p=.019) was positively and significantly correlated with the time they took to complete the tasks.

Final question vignettes: Direct robes of anxiety, comfort and perception

The final vignette questions aimed to directly evaluate participants' perception of anxiety, comfort and effectiveness of the think aloud interview (see Table 2 for descriptive statistics). In terms of participants' direct perception of feeling anxious with thinking aloud as they solved the four math-type tasks, 19.7% of students reported not feeling anxious, 72.7% reported experiencing mild to moderate anxiety, and 7.6% reported severe to extreme anxiety. In terms of participants' direct perception of comfort with thinking aloud, most of the participants stated that they felt comfortable (30.3% - very comfortable, 34.8% - comfortable, and 28.8% - moderately comfortable); and only 6.1% stated they did not feel comfortable voicing their thoughts while solving the math-type tasks. Finally, most of the participants (51.5%) reported they directly perceived the think-aloud interview as completely or very effective in measuring their cognitive processes and strategies for solving the math-type tasks. Thirty-four percent believed it was only moderately effective, whereas 13.6% considered the think aloud interview to be only slightly effective or not effective at all. Participants' directly reported anxiety and comfort with voicing their thoughts aloud were significantly correlated with their direct perception of the effectiveness of the think aloud process to gather information about their cognitive processes and strategies. For example, Spearman correlations showed a moderate but negative relationship between their direct perception of the effectiveness of the interview and their direct perception of anxiety (r_S =-.248, n=66, p=.045 <.05), indicating that students experiencing more anxiety perceived the think aloud process as being less effective. Likewise, there was a strong positive correlation between their direct perception of the effectiveness of the think aloud interview and their direct perception of comfort (r_S =.442, n=66, p<.001).

Motivation (SIMS)

Participants' reported motivation levels are shown in Table 3. Motivation was measured using the SIMS, which included four sub-scales: intrinsic motivation, identified regulation, external regulation, and amotivation. The scores on the intrinsic motivation and identified regulation were normally distributed. However, scores on the external regulation were negatively skewed and the scores on the amotivation were positively skewed. A positive Spearman correlation (r_s=.266, n=66, p=.031) was found between participants' identified regulation and their reported level of direct comfort during the think-aloud interview. Recall that identified regulation occurs when a behavior is perceived as chosen by oneself but the activity is not performed for itself but as the means to reach a final goal. Thus, this correlation indicates that students who report higher levels of perceived comfort also expressed more control with the reasons for engaging with the activity. In addition, external regulation exhibited a moderate but negative correlation with both participants' direct level of comfort level (r_s=-.246, n=66, p=.047) and perceived effectiveness of the think-aloud interview (r_s=-.370, n=66, p=.002). Again, external regulation focuses on behaviors a person feels obligated to engage in which are driven by rewards or to avoid negative consequences. Thus, this correlation indicates that students who perceived more comfort and effectiveness of the think-aloud interview were less likely to report feeling obligated to engage in the think-aloud interview. Finally, there was a moderate negative correlation (r_s=-.311, n=66, p<.024) between amotivation and students' direct level of comfort level with the think aloud interview.

Table 3
Means and Standard Deviations on the SIMS and ATMI Measures

	SIMS sub-scales				ATMI sub-scales		
	Intrinsic motivation	Identified regulation	External regulation	Amotivation	Affective	Behavioral	Cognitive
M	16.41	14.2	20.12	8.06	34.14	30.86	12.29
SD	5.491	5.054	5.696	3.378	10.661	4.117	3.63
Min.	4	5	6	4	15	19	7
Max.	28	28	28	17	54	38	21

Attitudes toward mistakes inventory (ATMI)

Participants' attitudes towards mistakes as measured by the ATMI's affective, behavioral, and cognitive sub-scales are presented in Table 3. Participants' scores ranged from 15 to 54 on the affective sub-scale, from 19 to 38 on the behavioral sub-scale, and from 7 to 21 on the cognitive sub-scale. The scores on the affective and behavioral sub-scales were normally distributed but the scores on cognitive sub-scale were positively skewed. As expected, a moderate positive correlation (r_s =.283, n=66, p=.021) was found between participants' behavioral attitudes toward mistakes and their accuracy on the four math-type tasks; meaning that students with better behavioral attitudes towards their mistakes scored better on the math-type tasks. Participants' emotional attitudes towards mistakes also shared a strong positive correlation (r_s=.477, n=65, p<.001) with their direct level of anxiety, indicating that students with more negative emotional attitudes towards their mistakes reported higher levels of perceived anxiety with the think aloud interview. There was also a moderate negative correlation between participants' emotional attitudes towards mistakes with their direct level of comfort level during the think-aloud interview (r_s=-.333, n=65, p=.007), indicating that more negative emotional attitudes were associated with less perceived comfort about the think aloud interview.

Regression Analyses

Based on the statistically significant correlations presented in the preliminary analyses, the following relationships were explored further: (a) predicting accuracy of performance on the math-type tasks based on participants' direct level of perceived comfort and behavioral attitudes towards mistakes (ATMI); b) predicting participants' indirect level of perceived effectiveness of the think-aloud interview based on their external regulation (SIMS) and time to complete the math-type tasks; and (c) predicting participants' direct level of perceived effectiveness of the think aloud interview based on their external regulation (SIMS) and direct level of perceived comfort.

First, an ordinal regression analysis was conducted to predict accuracy on the four math-type tasks based on participants' direct level of perceived comfort and behavioral attitudes towards mistakes. Ordinal regression was used instead of linear regression because total scores on math-type tasks showed a marked departure from normality, as it was bimodal in distribution. Ordinal regression does not require that the criterion variable display normality but it does require that the variable be ordered, meaning that smaller values indicate less of the variable than larger values. To meet the other assumptions for ordinal regression, the predictors (i.e., direct level of perceived comfort and behavioral attitudes towards mistakes) were dichotomized. The predictor variables were dichotomized because ordinal regression requires that the number of observations within each of the cells created by crossing each level of the predictors reflect a value greater than 5.

Participants' direct level of perceived comfort was dichotomized by assigning a value of 1 to responses indicating "bad" or "moderate" comfort, and a value of 2 to responses indicating "good" or "very good" comfort with the think-aloud interview. The behavioral attitudes towards

mistakes variable was dichotomized in two steps – first by charting the quartiles of the distribution and looking for natural cut point in scores. This step led to four levels: sub-scores below 28 (25% of the distribution), 29-31 (30% of the distribution), 32-34 (25% of the distribution), and sub-scores above 35 (20% of the distribution). Following the identification of quartiles and with inspection of the original distribution, the natural cut for dichotomizing the variable was at the 32-sub-score point. A value of 1 was assigned to scores at 31 or below and a value of 2 was assigned to scores at or above 32.

The SPSS ordinal regression procedure, which uses the Polytomous Universal Model or PLUM, was applied to the variables. The assumption of parallel lines for ordinal regression was met, -2 Log Likelihood difference = 35.57 – 32.44, Chi Square = 3.127, df=4, p=.537, indicating that the slope coefficient or relationship between the predictor variables and each category of the criterion variable (i.e., the logits) was the same. Shown in Table 4 are the parameter estimates.

Table 4
Summary of Ordinal Regression for Variables Predicting Accuracy

		Estimate	Std. Error	Wald	df	sig.	95% CI Lower Bound	95% CI Upper Bound
Threshold	Accuracy=1	-3.135	0.564	30.953	1	0	-4.24	-2.031
	Accuracy=2	-0.931	0.401	5.378	1	0.02	-1.717	-0.144
	Accuracy=3	0.08	0.385	0.043	1	0.836	-0.0674	0.834
Location	B_Attitudes=1	-1.074	0.477	5.07	1	0.024	-2.009	-0.139
	B_Attitudes=2	0[a]	.	.	0	.	.	.
	Direct Comfort=1	-1.077	0.504	4.565	1	0.033	-2.065	-0.089
	Direct Comfort=2	0[a]	.	.	0	.	.	.

Notes. Link function: Logit; "a" parameter set to zero because it is redundant.

The regression results can be interpreted as follows: For any category of average performance on the math-type tasks (i.e., 1 through 4), participants who scored at or less than 31 on the behavioral attitudes were less accurate in their performance than participants who scored

at or above 32. This difference was significant at p=.024. For the participants who were at or below 31 on behavioral attitudes, the expected log odds ratio decreased by 1.074 with each category of average performance. Likewise, for any category of average performance on the math-type tasks (i.e., 1 through 4), participants who rated their direct perceived comfort with the think aloud interview a 1 (i.e., bad or moderate comfort) were less accurate in their accuracy performance than those participants who rated it a 2 (i.e., good or very good). This difference was significant at p=.033. For the participants who rated their comfort a 1, the expected log odds ratio decreased by 1.077 with each category of average performance. In other words, participants who exhibited less favorable behavioral attitudes towards mistakes and who directly perceived the think-aloud interview as less comfortable were significantly more likely to be less accurate in their performance on the math-type tasks.

Fit indices showed the model to be a good fit to the data (i.e., meaning that the observed and expected counts in the cells were comparable): Pearson Goodness of fit, Chi-Square=6.907, df=7, p=.439; and deviance, Chi-Square=7.388, df=7, p=.390. Ordinal regression also leads to a type of R^2 measure, called the pseudo R-square, although these are not interpreted as straightforwardly as those for linear regression. Nonetheless, all three pseudo R-square measures were acceptable and indicated a moderate strength of association: Cox and Snell R^2=.154, Nagelkerke R^2=.166, and McFadden R^2=.064.

A second regression was calculated to predict participants' indirect perception of the effectiveness of the think aloud interview based on their external regulation (SIMS), time to complete the four math-type tasks, and their indirect level of perceived comfort. Linear regression was used for this analysis as the criterion variable (i.e., indirect perception of effectiveness) was approximately normally distributed. The initial analysis showed that the

assumptions were not violated. The collinearity statistics showed no multicollinearity (VIF values <10, Tolerance values >.2, see Table 5) and the Dubrin-Watson value of 2.152 was within the boundaries (not <1 or > 3) to meet the assumption of independent errors. As shown in Table 5, a significant step-wise regression equation was found, $F(3,62)=12.639$, $p<.000$, with R^2 of .379. In this case, inspection of the standardized regression coefficients indicated that only time ($p=.013 < .05$) and participants' indirect level of perceived comfort ($p=.000$) were statistically significant predictors of participants' indirect perception of the effectiveness of the think aloud interview. External regulation ($p=.515$) was not a significant predictor. The results of the regression equation can be interpreted as follows for the two significant predictors: For every standard deviation unit increase in time to complete the math-type tasks, there is an increase of .263 standard deviation units for participants' indirect level of perceived effectiveness of the think-aloud interview. Moreover, for every standard deviation unit increase in participants' indirect perceived comfort, there is an increase of .503 standard deviation units for their indirect level of perceived effectiveness of the think-aloud interview.

Table 5
Summary of Regression Analysis for Variables Predicting Indirect Effectiveness

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.616[a]	0.379	0.349	2.428	0.379	12.639	3	62	0.000	2.152

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	4.254	3.360		1.266	0.210	-2.463	10.971		
	external regulation	-0.038	0.059	-0.073	-0.654	0.515	-0.156	0.079	0.809	1.236
	time	0.003	0.001	0.263	2.551	0.013	0.001	0.005	0.939	1.065
	indirect comfort	0.607	0.130	0.503	4.651	0.000	0.346	0.867	0.857	1.166

Predictors: (Constant), external regulation, time, indirect comfort

Criterion: indirect effectiveness

A third linear regression was conducted to predict participants' direct perception of the effectiveness of the interview based on their external regulation (SIMS) and direct perception of comfort with the think-aloud interview. The initial analysis showed that the assumptions were met. The collinearity statistics showed no multicollinearity (VIF=1.066 <10, Tolerance=.938 >.2) and the Dubrin-Watson value of 2.196 was within the boundaries (not <1 or > 3) to meet the assumption of independent errors. As shown in Table 6, a significant regression equation was found, $F(2,63)=10.558$, $p=.000$, with R^2 of .251. The regression equation can be interpreted as follows: For every standard deviation unit increase in participants' external regulation score, there was a decrease of -.229 standard deviation units on their direct perception of the effectiveness of the interview. Likewise, for every standard deviation unit increase in their direct perception of comfort with the think-aloud interview, there was an increase of .391 standard deviation units on their direct perception of the effectiveness of the interview. Inspection of the standardized coefficients for both predictors were examined and both external regulation

(p=.045<.05) and direct comfort (p=.001) show statistical significance in predicting participants' direct perception of the effectiveness of the interview. These results are elaborated in the discussion section in light of the research questions guiding the study.

Table 6
Summary of Regression Analysis for Variables Predicting Direct Effectiveness

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.501[a]	0.251	0.227	0.834	0.251	10.558	2	63	0.000	2.196

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	2.654	0.667		3.978	0.000	1.321	3.987		
	external regulation	-0.038	0.019	-0.229	-2.037	0.046	-0.076	-0.001	0.938	1.066
	direct comfort	0.407	0.117	0.392	3.482	0.001	0.173	0.640	0.938	1.066

Predictors: (Constant), external regulation, direct comfort

Criterion: direct effectiveness

Discussion

The objective of this study was to evaluate whether interviewee variables such as individual cognitive ability, motivation, attitudes towards mistakes and perception of think aloud interviews influenced accuracy of performance during the interview. There is limited research on the human factors that might impact the accuracy of think-aloud interview data. The current research attempted to evaluate the factors that might influence the accuracy and quality of response process data during think-aloud interviews, which could introduce bias into the interpretation of verbal reports.

The results of the present study indicate that situational motivation and cognitive abilities of interviewees were not associated with their performance on the math-type tasks during the

think-aloud interview. Given the uniqueness of the study in terms of focusing on think aloud interviews specifically, it is not surprising that the current study did not confirm previous research showing the effect of intrinsic and extrinsic motivation (e.g., Lin, McKeachie, & Kim, 2003; Mauro, 2007) on students' task performance. Several reasons may account for this. First, the domain of the study was different from previous work as it was focused on participants' motivation during the think aloud interview. The author assessed situational motivation pertaining specifically to an activity rather than general achievement. Second, it is possible that there were not enough math-type tasks to distinguish performance levels in light of motivation. Unlike previous research that has shown a significant relationship between cognitive abilities and students' school performance/ achievement (e.g., Goldstein, Princiotta, & Naglieri, 2015; Ridgell, & Lounsbury, 2004), the current study did not find strong association of participants' math-type task performance and their cognitive abilities. Again, the lack of an association might be due to the limited number of math-type tasks. However, the focus of the study was not to replicate well-established findings showing a relationship between IQ and achievement. But, rather, to explore whether cognitive ability might influence think aloud interview performance. The results suggest that cognitive ability does not as it was not correlated with any of the key variables – accuracy, motivation, attitudes towards mistakes and perception of the think aloud interview. The next level of analysis requires investigation into whether participants' cognitive abilities influence the content of their verbal reports.

Although no effects of cognitive ability were found, participants' direct level of perceived comfort with thinking aloud and behavioral attitudes towards mistakes were significantly related to the accuracy of their performance on the math-type tasks. Students who reported feeling more comfortable during the think-aloud interview performed more accurately

on the math-type tasks than those students who reported less comfort. This result suggests that participants who do not feel comfortable thinking aloud might experience some type of cognitive interference arising from emotion or distraction, which could lead to disrupted performance. Students who reported feeling less comfortable also reported experiencing greater anxiety during the think-aloud interview, which lends support to the idea that there is some disruption to performance as anxiety is known to disrupt performance on tasks of moderate complexity (Ashcraft & Kirk, 2001). This needs to be investigated further. There are implications for this result on studies employing think-aloud interviews. For students who do not feel comfortable with voicing their thoughts aloud, additional supports may be required such as more time might be required for practice before initiating the actual interview.

In addition, students who had higher scores on behavioral attitudes towards mistakes were found to be more accurate in their solutions of the math-type tasks. Interestingly, participants who are likely to take remediate action when a mistake is made performed more accurately on the tasks under think-aloud interview conditions than those participants who were less likely to take remediate action. One way to explain this result is that students who are in the habit of working through their mistakes might have more practice working through their problem solving or may be open to experience, including feedback that is known to lead to more meaningful learning (Chamorro-Premuzic & Furnham, 2006).

The current study also investigated interviewees' perception of the interview in terms of its effectiveness as a tool with which to measure students' response processes and strategies. The results indicated that students' indirect perception of the effectiveness of the think-aloud interview was positively associated with their indirect levels of comfort and time to complete the math-type tasks. Although these findings are correlational, it is probably the case that students

who feel more comfortable expressing their thoughts tend to view the interview as an efficacious method of tapping into response processes, and therefore take the time to engage with the interview.

Time to complete the tasks was not associated with students' direct perception of anxiety, comfort or the effectiveness of the think-aloud interview. However, time was associated with participants' indirect perception of effectiveness of the interview – students who took more time also reported higher ratings of indirect perceived effectiveness. The lack of an effect of time for the direct measure of perception may be due to some bias in the direct measure. Using the direct measure of perception, students may have felt some incentive to present a socially desirable self by indicating the think-aloud interview was effective in measuring the processes and strategies even when they quickly responded to all math-type tasks. Students' direct perception of the effectiveness of the interview was also related to their external motivation. Students who were more externally motivated were less inclined to perceive think-aloud interview as an effective to measure of their response processes and strategies. In addition, students with various levels of anxiety did not significantly differ in their perception of the interview.

Limitations

Several limitations of this study should be noted. Participants` level of comfort could vary depending on the type of the math task and could increase or decrease during the process of solving the tasks as existing research indicates that task difficulty can influence task accuracy (Ericsson, & Simon, 1993). Participants' perception of effectiveness of the think-aloud interview might vary not only based on their level of comfort and external motivation but it might as well depend on other factors (e.g., the type of task, perception of whether they solved the task correctly). In fact, in this study, a statistically significant negative correlation was found between

participants' indirect perception of anxiety and their indirect perception of effectiveness of the interview (r=-.307, p=.012). However, when this measure of anxiety was entered into the regression equation to predict perception of effectiveness, it was not found to be a statistically significant predictor. This is likely the case because this measure of anxiety was also statistically associated with indirect perception of comfort (r=.623, p<.000), the latter of which was a strong predictor of indirect perception of effectiveness. Thus, more sophisticated statistical analysis (e.g., structural equation modeling) with larger samples may be needed to evaluate any mediation or moderation effects between anxiety and comfort measures in relation to indirect levels of perceived effectiveness.

The results of the study can also be enhanced by examining the retrospective reports gathered after participants complete each math-type task to investigate cognitive factors more precisely. This analysis is going to be the next step as these data have been collected. In addition, the study should be replicated with various populations to examine the reliability of generalizing the results to other student populations.

Conclusions

The results of the study suggest that ensuring participants feel comfortable (or are less anxious) with think-aloud interview may be critical to ensure the accuracy of the tasks solutions is not affected during the process and their perception of the interview is constructive. Researchers can further explore what contributes to participants' level of comfort under think-aloud interview conditions. The qualitative data gathered during the think aloud interview may provide insight on factors that contribute to participants' comfort level during their problem solving on the math-type tasks. One of the important factors that can impact level of comfort is the trust a student is able to establish with the person conducting interview. Moreover, the

examination of the retrospective reports can be explored to see whether the perception of task difficulty impacted participants' level of comfort and therefore the perception of the think-aloud interview effectiveness. In addition, only the students who never took part in think-aloud interview studies were invited to participate in the study. Further research can investigate whether people who have experience with think-aloud aloud interview studies feel more comfortable when participating in such studies again, and whether they differ in their comfort level as compared to those who take part in the study for the first time.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Ashcraft, M. H., & Kirk, E. P. (2001). The relationships among working memory, math anxiety, and performance. *Journal of Experimental Psychology: General, 130*, 224-237.

Bruckner, S., & Pellegrino, J. W. (2016). Integrating the analysis of mental operations into multilevel models to validate an assessment of higher education students` competency in business and economics. *Journal of Educational Measurement*, 53 (3), 293-311.

Carroll, J. B. (1997). The three-stratum theory of cognitive abilities. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp.122-130). New York: Guilford.

Cattell, R. B. (1971). *Abilities: Their structure, growth, and action*. Boston: Houghton Mifflin.

Ceci, S. J. (1993). Contextual trends in intellectual development. *Developmental Review*, 13, 403-435.

Chamorro-Premuzic, T., & Furnham, A. (2006). Intellectual competence and the intelligent personality: A third way in differential psychology. *Review of General Psychology, 10*, 251–267.

Church, M. A., Elliot, A. J., & Gable, S. L. (2001). Perception of classroom environment, achievement goals, and achievement outcomes. *Journal of Educational Psychology*, 93 (1), 43-54.

Covington, M. V. (2000). Goal theory, motivation and school achievement: An integrative review. *Annual Review of Psychology*, 51, 171–200.

Deci, E. L., Koestner, R., & Ryan, R. M. (2001). Extrinsic rewards and intrinsic motivation in education: Reconsidered once again. *Review of Educational Research*, 71, 1–27.

Deci, E. L, & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior.* New York: Plenum.

DordiNejad, F. G., Hakimi, H., Ashouri, M., Dehghani, M., Zeinali, Z., Daghighi, M. S., & Bahrami, N. (2011). On the relationship between test anxiety and academic performance. *Procedia Social and Behavioral Sciences*, 15, 3774-3778

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

Evans, J. J., Floyd, R. G., McGrew, K. S., & Leforgee, M. H. (2002). The relations between measures of Cattell-Horn-Carroll (CHC) cognitive abilities and reading achievement during childhood and adolescence. *School Psychology Review*, 31, 246-262.

Farsides, T., $ Woodfield, R. (2003). Individual differences and undergraduate academic success: the roles of personality, intelligence, and application. *Personality and Individual differences*, 34, 1225-1243.

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, 137 (2), 316-344.

Gierl, M. J. (1997). Comparing the cognitive representations of test developers and students on mathematical achievement test using Bloom`s taxonomy. *Journal of Educational Research*, 91, 34-44.

Gilhooly, K. J., Fioratou, E., & Henretty, N. (2010). Verbalization and problem solving: Insight and spatial factors. British Journal of Psychology, 101, 81-93.

Goldstein, S., Naglieri, J. A., & Princiotta, D. (2015). *Handbook of intelligence: evolutionary theory, historical perspective, and current concepts.* New York: Springer.

Guay, F., Vallerand, R. J., & Blanchard, C. (2000). On the assessment of situational intrinsic and extrinsic motivation: The situational motivation scale (SIMS). *Motivation and Emotion*, 24, 175-213.

Guilford, J. B. (1988). Some changes in the structure of the intellect model. *Educational and Psychological Measurement*, 48, 1-4.

Hong, E. (1999). Test anxiety, perceived test difficulty, and test performance: Temporal patterns of their effects. *Learning and Individual Differences*, *11(4),* 431-447.

Horn, J. L. (1985). Remodeling old models of intelligence: G-Gc theory. In B. B. Wolman (Ed.), *Handbook of intelligence* (pp.267-300). New York: Wiley.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50, 1-73.

Katz, I. R., Bennett, R. E., & Berger, A. E. (2000). Effects of response format on difficulty of SAT-mathematics items: It`s not the strategy. *Journal of Educational Measurement*, 37 (1), 39-57.

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 misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, 23, 6-15.

Leighton, J. P. (2013). Item difficulty and interviewer knowledge effects on the accuracy and consistency of examinee response processes in verbal reports. *Applied Measurement in Education*, 26, 136-157.

Leighton, J. P. (2017). *Using think-aloud interviews and cognitive labs in educational research.* New York, NY: Oxford University Press.

Leighton, J. P., Rogers, W. T., & Maguire, T. O. (1999). Assessment of student problem solving on ill-defined tasks. *Alberta Journal of Educational Research*, 45, 409-427.

Leighton, J. P., Tang, W., & Guo, Q.. (2018). Undergraduate students` attitudes towards mistakes in learning and academic achievement. *Assessment & Evaluation in Higher Education,* 43, 612-628.

Lina, Y. G., McKeachieb, W. J., & Kim, Y. C. (2003). College student intrinsic and/or extrinsic motivation and learning. *Learning and Individual Differences*, 13, 251-258.

Mauro, M. R. (2007). Predictors of success in college and career: Effects of personality, motivation, and drive (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3613810)

Morse, D. T., & Morse, L. W. (2002). Are undergraduate examinees` perceptions of item difficulty related to item characteristics? *Perceptual and Motor Skills,* 95, 1281-1286.

Murray C., & Wren, C. T. (2003). Cognitive, academic, and attitudinal predictors of the grade point averages of college students with learning disabilities. *Journal of Learning Disabilities*, 36, 407-415.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231-259.

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

Pekrun, R. Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist, 37*, 91–105.

Piaget, J. (1972). *Psychology of intelligence*. Totowa, NJ: Littlefield Adams.

Reber, A. S., Allen, R., & Reber, E. Sarah. (2009). *The Penguin dictionary of psychology.* 4th ed. London: Penguin books.

Reynolds, C. R., & Hickman, J. A. (2004). Draw-A-Person Intellectual Ability Test for Children, Adolescents, and Adults (DAP:IQ) examiner`s manual. Austin, TX: Pro-ed.

Ridgell, S. D., & Lounsbury, J. W. (2004). Predicting academic success: General intelligence, "Big Five" personality traits, and work drive. *College Student Journal*, 38, 607-619.

Ruthig, J. C., Perry, R. P., Hladkyj, S., Hall, N. C., Pekrun, R., & Chipperfield, J. G. (2008). Perceived control and emotions: interactive effects on performance in achievement settings. Social Psychological Education, 11, 161-180.

Ryan, R. M. (2012). *The Oxford handbook of human motivation.* New York: Oxford University Press.

Wang, C., & Gierl, M. J. (2011). Using the attribute hierarchy method to make diagnostic inferences about examinees` cognitive skills in critical reading. *Journal of Educational Measurement*, 48 (2), 165-187.

Whiston, S. C. (2009). *Principles and applications of assessment in counseling.* 3rd ed. Australia: Brooks/Cole Cengage Learning.

Zeidner, M. (2007). Test anxiety in educational contexts: Concepts, findings, and future directions. In P. A. Schutz & R. Pekrun (Eds.), *Emotion in education* (pp. 165–184). San Diego, CA: Academic Press.

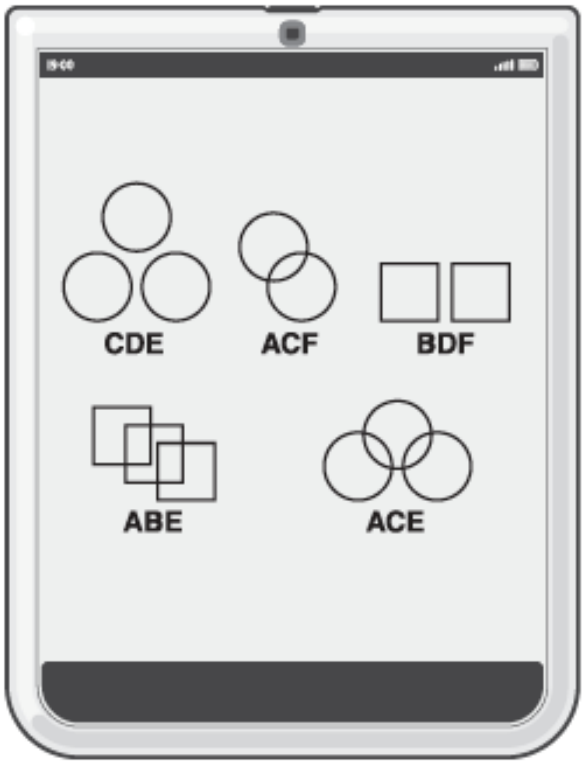
Appendix 1

Four Examples of Publicly Released Academic Tasks: Diploma Practice Tasks

Math-Type Task 1

Use the following information to answer the next question.

In a particular game, drawings consist of circles or squares. Different letters are used to describe the type of shape, the number of shapes, and the arrangement of the shapes in the drawing. Five drawings from a particular game are shown below.



2. Based on the drawings above, which of the following statements about the letter *A* is true?

A. The letter *A* represents circles.
B. The letter *A* represents squares.
C. The letter *A* represents overlapping shapes.
D. The letter *A* represents the number of shapes in the drawing.

Math-Type Task 2

Use the following information to answer the next question.

There are 35 students in John's homeroom class. There are 5 students who take English and biology, and 7 students who take neither of these subjects. There are 3 more students taking only English than there are students taking only biology.

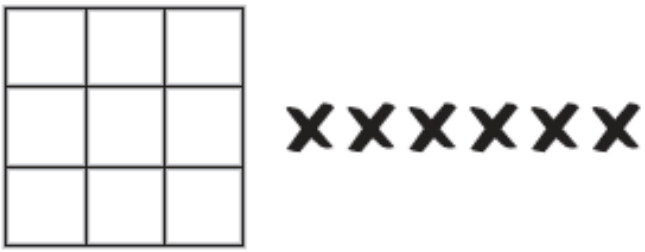
12. The number of students in John's homeroom who take biology only is

A. 10
B. 13
C. 15
D. 20

Math-Type Task 3

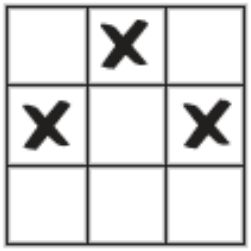
Use the following information to answer the next question.

To solve a particular logic puzzle, an X must be placed in six of the squares in the grid shown below so that no three Xs lie in a vertical, horizontal, or diagonal line.

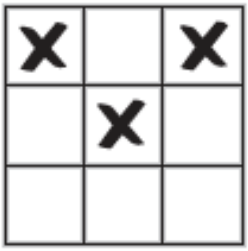


14. Which of the following opening placements for the first three Xs could lead to a successful solution for this logic puzzle?

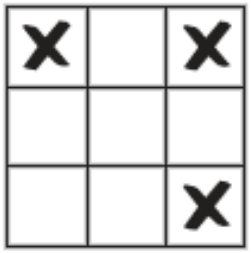
A.



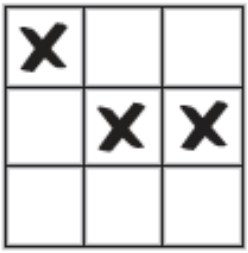
B.



C.



D.



Math-Type Task 4

Use the following information to answer the next question.

A candy dish contains 2 blue candies, 1 red candy, 3 white candies, and 2 green candies. Two candies will be randomly selected from the dish, one after the other, without replacement.

14. The probability of selecting a blue candy and then a candy that is **not** blue is

A. $\frac{3}{14}$

B. $\frac{3}{16}$

C. $\frac{1}{16}$

D. $\frac{1}{28}$

The Situational Motivation Scale

DIRECTIONS: Read each item carefully. Using the scale below, **please circle the number that best describes the reason why you are currently engaged in this activity.**

Answer each item according to the following scale:
1: corresponds not at all; 2: corresponds to very little; 3: corresponds to a little; 4: correspond to moderately; 5: corresponds to enough 6: corresponds to a lot; 7: corresponds to exactly

Why are you currently engaged in this activity?

1. Because I think that this activity is interesting.	1 Not at all	2 Very Little	3 A little	4 Moderately	5 Enough	6 A lot	7 Exactly
2. Because I am doing it for my own good.	1 Not at all	2 Very Little	3 A little	4 Moderately	5 Enough	6 A lot	7 Exactly
3. Because I am supposed to do it.	1 Not at all	2 Very Little	3 A little	4 Moderately	5 Enough	6 A lot	7 Exactly
4. There may be good reasons to do this activity, but personally I don't see any.	1 Not at all	2 Very Little	3 A little	4 Moderately	5 Enough	6 A lot	7 Exactly
5. Because I think that this activity is pleasant.	1 Not at all	2 Very Little	3 A little	4 Moderately	5 Enough	6 A lot	7 Exactly
6. Because I think that this activity is good for me.	1 Not at all	2 Very Little	3 A little	4 Moderately	5 Enough	6 A lot	7 Exactly
7. Because it is something that I have to do.	1 Not at all	2 Very Little	3 A little	4 Moderately	5 Enough	6 A lot	7 Exactly
8. I do this activity but I am not sure if it is worth it.	1 Not at all	2 Very Little	3 A little	4 Moderately	5 Enough	6 A lot	7 Exactly
9. Because this activity is fun.	1 Not at all	2 Very Little	3 A little	4 Moderately	5 Enough	6 A lot	7 Exactly
10. By personal decision.	1 Not at all	2 Very Little	3 A little	4 Moderately	5 Enough	6 A lot	7 Exactly
11. Because I don't have any choice.	1 Not at all	2 Very Little	3 A little	4 Moderately	5 Enough	6 A lot	7 Exactly
12. I don't know; I don't see what this activity brings me/	1 Not at all	2 Very Little	3 A little	4 Moderately	5 Enough	6 A lot	7 Exactly
13. Because I feel good when doing this activity.	1 Not at all	2 Very Little	3 A little	4 Moderately	5 Enough	6 A lot	7 Exactly
14. Because I believe that this activity is important for me.	1 Not at all	2 Very Little	3 A little	4 Moderately	5 Enough	6 A lot	7 Exactly
15. Because I feel that I have to do it.	1 Not at all	2 Very Little	3 A little	4 Moderately	5 Enough	6 A lot	7 Exactly
16. I do this activity, but I am not sure it is a good thing to pursue it.	1 Not at all	2 Very Little	3 A little	4 Moderately	5 Enough	6 A lot	7 Exactly

Attitudes Towards Mistakes Inventory

DIRECTIONS: Please read each item carefully. Using the scale below, please respond to the following items as honestly as you can. When responding to each of the items, please relate them to your academic experience in general.

	Strongly Disagree 1	Disagree 2	Neither Disagree nor Agree 3	Agree 4	Strongly Agree 5
1. When I make mistakes in group discussions, I am afraid that others look down upon me.	1	2	3	4	5
2. When I make mistakes on my assignment, I am quite curious about where I went wrong.	1	2	3	4	5
3. I believe successful students make fewer mistakes during learning than others.	1	2	3	4	5
4. If I make mistakes in group discussions, I don't want others to notice them.	1	2	3	4	5
5. When I make mistakes on an exam, I feel motivated to study harder.	1	2	3	4	5
6. I believe it is smart to avoid making mistakes during learning.	1	2	3	4	5
7. When I make mistakes answering classroom questions, I am overwhelmed with embarrassment.	1	2	3	4	5
8. If I make mistakes on my assignment, I will redo it.	1	2	3	4	5
9. I believe making mistakes is not an efficient way to learn academic materials.	1	2	3	4	5
10. I seldom feel bothered by the mistakes I make in group discussions.	1	2	3	4	5
11. When I see mistakes on an exam, I meet the instructor to review the errors.	1	2	3	4	5
12. I believe I do not learn much from making mistakes in learning.	1	2	3	4	5
13. When I make mistakes answering classroom questions, I still feel confident about my ability.	1	2	3	4	5
14. When I make mistakes on an exam, I find similar exercises to practice.	1	2	3	4	5
15. I believe I gain knowledge from making mistakes.	1	2	3	4	5
16. When I make mistakes in class, I worry that other students may laugh at me.	1	2	3	4	5
17. When I make mistakes on my assignment, I try to find out why by checking the class notes.	1	2	3	4	5
18. I believe making mistakes is a necessary part of learning.	1	2	3	4	5
19. When I make mistakes answering classroom questions, I become anxious.	1	2	3	4	5
20. When I make mistakes on my assignment, I compare my answers with the examples in the class notes.	1	2	3	4	5
21. I make mistakes on my assignment because I am not smart enough.	1	2	3	4	5
22. When I make mistakes answering classroom questions, I feel humiliated.	1	2	3	4	5
23. When I make mistakes on my assignment, I review what was discussed in class.	1	2	3	4	5
24. When I make mistakes answering classroom questions, I am mortified.	1	2	3	4	5
25. When I make mistakes answering classroom questions, I am not disappointed with my answer.	1	2	3	4	5
26. I feel safe in group discussions even if I make mistakes.	1	2	3	4	5

Anchoring Vignettes Activity

Instructions: In the following pages, you will read 6 short stories (vignettes) about individuals who have taken part in think-aloud interviews of some kind. Even though the stories might appear to be similar, there are differences. Please read the stories carefully and then respond to the questions that follow as best you can given your impression of the stories.

Vignette #1

Nick had always struggled with math class in high school. It took him awhile to solve math problems but in the end, he earned a good final grade. Now in university, when his math professor asked him to think aloud as he was solving math problems, he seemed to be very concentrated and kept voicing his thoughts consistently.

Given the above information, please put yourself in Nick's position and answer the following questions as best you can:

How much difficulty do you think Nick experienced with feeling anxious, uneasy, or worried when he was voicing his thoughts out loud during the solving of math problems?

1. None
2. Mild
3. Moderate
4. Severe
5. Extreme

How would you rate Nick's experience with being comfortable in voicing his thoughts out loud when solving the math problems?

1. Very good
2. Good
3. Moderate
4. Bad
5. Very bad

How effective might Nick consider the think aloud process to be in gathering information about a student's cognitive processes (or strategies) during the solving of math problems?

1. Completely effective
2. Very effective
3. Moderately effective
4. Slightly effective
5. Not effective at all

Vignette #2

Ann had excellent skills in solving any math problem in school as she used to take part in various math problem solving competitions. However, now in university, she was not sure what was expected of her during a think aloud interview for which she signed up. Therefore, she kept going back and forth between voicing her immediate thoughts and commenting on how she solved a math task after she solved it.

Given the above information, please put yourself in Ann's position and answer the following questions as best you can:

How much difficulty do you think Ann experienced with feeling anxious, uneasy, or worried when she was voicing her thoughts out loud during the solving of math problems?

1. None
2. Mild
3. Moderate
4. Severe
5. Extreme

How would you rate Ann's experience with being comfortable in voicing her thoughts out loud when solving the math problems?

1. Very good
2. Good
3. Moderate
4. Bad
5. Very bad

How effective might Ann consider the think aloud process to be in gathering information about a student's cognitive processes (or strategies) during the solving of math problems?

1. Completely effective
2. Very effective
3. Moderately effective
4. Slightly effective
5. Not effective at all

Vignette #3

Tasha put a lot of efforts to complete all her math homework and come prepared to school. She did not like group work in class and never went out with her friends for after school activities. Math was not her strongest area of expertise. During the think aloud interview with her math coach she voiced her thoughts constantly.

Given the above information, please put yourself in Tasha's position and answer the following questions as best you can:

How much difficulty do you think Tasha experienced with feeling anxious, uneasy, or worried when she was voicing her thoughts out loud during the solving of math problems?

1. None
2. Mild
3. Moderate
4. Severe
5. Extreme

How would you rate Tasha's experience with being comfortable in voicing her thoughts out loud when solving the math problems?

1. Very good
2. Good
3. Moderate
4. Bad
5. Very bad

How effective might Tasha consider the think aloud process to be in gathering information about a student's cognitive processes (or strategies) during the solving of math problems?

1. Completely effective
2. Very effective
3. Moderately effective
4. Slightly effective
5. Not effective at all

Vignette #4

Vince loved being engaged in social projects on campus. He enjoyed challenging tasks and wanted to succeed in everything he did. Vince did not think he solved all math problem tasks correctly during the think aloud study he participated in but he kept voicing his thoughts steadily throughout the process.

Given the above information, please put yourself in Vince's position and answer the following questions as best you can:

How much difficulty do you think Vince experienced with feeling anxious, uneasy, or worried when he was voicing his thoughts out loud during the solving of math problems?

1. None
2. Mild
3. Moderate
4. Severe
5. Extreme

How would you rate Vince's experience with being comfortable in voicing his thoughts out loud when solving the math problems?

1. Very good
2. Good
3. Moderate
4. Bad
5. Very bad

How effective might Vince consider the think aloud process to be in gathering information about a student's cognitive processes (or strategies) during the solving of math problems?

1. Completely effective
2. Very effective
3. Moderately effective
4. Slightly effective
5. Not effective at all

Vignette #5

Emma took a University Math class in high school and graduated with 4.1 GPA. She experienced anxiety whenever she was asked to talk in class. Emma completed a study where the researchers used the think aloud interview with a few reminders to Emma to "keep talking."

Given the above information, please put yourself in Emma's position and answer the following questions as best you can:

How much difficulty do you think Emma experienced with feeling anxious, uneasy, or worried when she was voicing her thoughts out loud during the solving of math problems?

1. None
2. Mild
3. Moderate
4. Severe
5. Extreme

How would you rate Emma's experience with being comfortable in voicing her thoughts out loud when solving the math problems?

1. Very good
2. Good
3. Moderate
4. Bad
5. Very bad

How effective might Emma consider the think aloud process to be in gathering information about a student's cognitive processes (or strategies) during the solving of math problems?

1. Completely effective
2. Very effective
3. Moderately effective
4. Slightly effective
5. Not effective at all

Vignette #6

Ed had always disliked math class in school and he enjoyed debating his teachers about why math class was not important for his future career as a comedian. During think aloud interview in which he participated in, Ed commented repeatedly a few times on each step he took to solve the math tasks.

Given the above information, please put yourself in Ed's position and answer the following questions as best you can:

How much difficulty do you think Ed experienced with feeling anxious, uneasy, or worried when he was voicing her thoughts out loud during the solving of math problems?

1. None
2. Mild
3. Moderate
4. Severe
5. Extreme

How would you rate Ed's experience with being comfortable in voicing his thoughts out loud when solving the math problems?

1. Very good
2. Good
3. Moderate
4. Bad
5. Very bad

How effective might Ed consider the think aloud process to be in gathering information about a student's cognitive processes (or strategies) during the solving of math problems?

1. Completely effective
2. Very effective
3. Moderately effective
4. Slightly effective
5. Not effective at all

Final Questions

How much difficulty did YOU have with feeling anxious, unease, or worried when you were voicing your thoughts out loud during the solving of math problems?

1. None
2. Mild
3. Moderate
4. Severe
5. Extreme

How would you rate YOUR experience with being comfortable to voice your thoughts out loud when solving the math problems?

1. Very good
2. Good
3. Moderate
4. Bad
5. Very bad

How effective do YOU consider the think aloud interview in gathering information on your cognitive processes (or strategies?) during the solving of math problems?

1. Completely effective
2. Very effective
3. Moderately effective
4. Slightly effective
5. Not effective at all

THANK YOU!

Demographic Information

Date:_____

Participant ID_____

Gender (Please circle one): Female / Male / Prefer not to disclose

Date of Birth (mm/dd/yy):_____/_____/_____

Ethnicity: _____

Major (and Minor of available): _____

Year: First / Second / Third / Fourth

High School GPA: _____

Last Year`s University GPA: _____

Appendix 2

Think Aloud Interview Script 2017/ 2018

Introduction

Hi, my name is _____ and I want to thank you for participating in this study. Before we start, please read over this consent form and sign if you agree to participate in this study. As the consent form indicates, please keep our discussions today confidential so as to not compromise the results of the study. This study is scheduled for approximately 2 hours. Please let me know at any time if you need a restroom break and we can stop and continue. This study involves thinking aloud, which essentially means voicing your thoughts out loud as you try to solve different tasks. This is a method that is under study because it is often used as a way to double check that test items such as those found on achievement tests are probing certain knowledge and/or skills in examinees. Please know that all the information you provide today is confidential and all data becomes anonymous before it is analyzed.

I will be showing you 4 math-type tasks today and I would like you to tell me all your thoughts as you try to solve each one. I am not a math expert so I will not be able to evaluate the correctness or incorrectness of your thoughts or methods for solving the tasks. I say this only because sometimes participants will ask me if they are "on the right track" and I won't be able to say one way or the other. If you go silent for more than 10 seconds while you're solving a task, I will remind you to keep talking. Also, I would like you to speak loudly because I want to make sure the audio-recorder captures your thoughts. Often people begin to whisper and we lose important information. After you finish voicing your thoughts for a task, I will ask you some questions about how you solved the task. We will do some practice problems to get us started and comfortable with the process. But, before we get to the practice problems, let me finish telling you what we will be doing today. So, after we finish with the math-type tasks, I will ask you to complete some additional activities – you do not have to think aloud when completing these additional activities. Do you have any questions at this stage?

Please note that I will be recording our interview today and taking notes so that we have as complete a confidential record as possible of the discussion. Okay, so let's go to discussing what it means to think aloud and the practice questions.

Process of Thinking Aloud and Practice Phase

I want to familiarize you with the procedure of thinking aloud. What I mean by think aloud is that I want you to say your thoughts out loud from the moment you finish hearing a practice question or task to your final answer – so please keep voicing your thoughts throughout the moment you hear the question, through the process of solving it until you tell me your final answer. I would like you to talk aloud as much as you comfortably can during that time. Don't try to plan or explain what you say. Just act as if you are alone and speaking to yourself. I know this seems awkward but please try to keep talking while you are coming up with the answer to each task and please remember to project your voice so that the recorder picks up your voice. If you are silent for a long time, I'll remind you to keep talking. Likewise, if you are speaking too softly, I will ask you to speak louder. Do you understand what I would like you to do? We will begin with a practice question. First, listen to the question, then answer it as soon as you can. Are you ready?

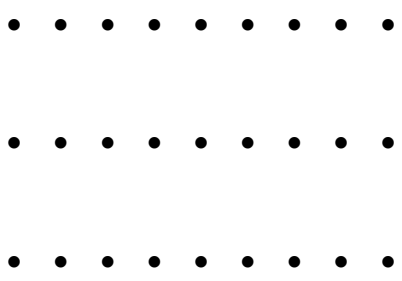
[*Practice Problem #1: What is the sixth letter after B? – see materials*]

Thank you. Did you have any other thoughts as you came up with the answer to this question? I want you to think those thoughts out loud as they occur to you.

Let us review this practice problem briefly. Chances are that the letter "H" didn't immediately occur to you after hearing the question. You probably had to go through several steps to find the answer. Had you summarized your thinking during this question rather than reporting the sequence of actual thoughts aloud, you might have said that you found the letter H by counting through the alphabet. But, when people actually solve this problem out loud, they usually say a sequence of individual letters, such as B, then C, D, E, F, and G, before the answer H. Because I am interested in knowing the thoughts you had as you answered the question, please voice your thoughts as you are solving the problem instead of summarizing your thoughts. Also, please remember that there is no need to explain your thoughts**.** Just say what you are thinking--even if it doesn't always seem grammatical or you're afraid that it won't make sense.

Let's do another question. Are you ready? I'm going to show you a dot grid and ask you to tell me how many dots in the grid.

(Practice Problem #2: "The 27-dot grid." Ask "How many dots are there"? – see materials)



Thank you. Can you recall any other thoughts? Please remember that there is no need to explain your thoughts to someone else**.** Just say what you are thinking--even if it doesn't always seem grammatical or you're afraid that it won't make sense.

Now let's begin the actual think aloud interview. Please take your time. We are here for two hours and one hour needs to be devoted to the tasks. This is not a test. We really want you to solve the task as you would on your own. You will be presented with four math-type problems – one at a time. Please tell me all that you are thinking from the moment you see the task until you finish solving it. Then I will ask you some follow-up questions before we move to the next task. If you finish the tasks early I will give you the additional ones to solve. Do you have any questions? OK. **I will now turn on the audio recorder** [*indicate day and time and participant ID*].

Math-Type Tasks (Think Aloud Phase)

Here is the booklet with the math-type tasks [*place booklet in front of participant*]. Please remember that I am not analyzing your thoughts or the strategies you use to solve the tasks while you voicing your thoughts. I will remind you to please keep talking as you solve each task. If you are silent for more than 10 seconds, I will ask you to "please keep talking." We will do one task at a time so please tell me once you are done solving the task – you may write on the paper if you wish and please do not forget to indicate your final answer on the sheet. After you have complete the task, I will ask you a series of five questions about how you solved the task. Then we will move on to the next task. Do you have any questions? Okay, let's start.

Post-Think Aloud Questions for Each Task

Okay, now that you are finished with the task, I will ask you the five questions….

1. Can you tell me the steps or any strategies you used to solve the task?
2. Did you find the task difficult or easy or so-so?
3. Did you feel any emotions as you were solving the task?
4. Is there anything you would change about the task?
5. How difficult or easy was it to solve the task while voicing your thoughts?

Ok, thank you, let's go to the next task.

Survey Activities

Now that we are finished with the math-type tasks, I will ask you to complete a series of activities. **I will turn off the audio-recorder as we are not thinking aloud for this phase of the study.** We will begin with an activity involving drawing a person which is a measure of the way people process information. Then I will ask you to complete scales designed to measure motivation, including attitudes about mistakes, your responses to how you think others might view think-aloud interviews and finally a short survey about yourself. Please take your time, read each activity carefully and answer all questions. Do you have any questions?

Appendix 3

GPA Scores Distribution

Letter Grade	Percent Grade	4.0 Scale
A+	97-100	4
A	93-96	4
A-	90-92	3.7
B+	87-89	3.3
B	83-86	3
B-	80-82	2.7
C+	77-79	2.3
C	73-76	2
C-	70-72	1.7
D+	67-69	1.3
D	65-66	1
E/F	Below 65	0

https://pages.collegeboard.org/how-to-convert-gpa-4.0-scale