USING THINK ALOUD PROTOCOLS IN THE VALIDITY INVESTIGATION OF AN ASSESSMENT OF COMPLEX THINKING

by

Juliette Lyons-Thomas

B.Sc., McGill University, 2006
M.A., New York University, 2008

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies

(Measurement, Evaluation, and Research Methodology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2014

Abstract

Validation requires the collection of evidence that supports inferences from an assessment. Think aloud protocols (TAPs) are one method of collecting validity evidence for assessments; however this technique is rarely used in validity investigations of complex thinking. The first research question investigated the use of TAPs as a method of validation for assessments of complex thinking. Specifically, the research explored how TAPs add to validity investigations beyond the information that psychometric analyses provide. TAPs were collected from 35 students using a historical thinking measure. A large-scale administration of the same assessment to 441 students provided data for the psychometric analyses. The TAP data were coded and compared to the psychometric data in order to investigate the first research question. It was found that TAPs are valuable by providing information that is consistent with psychometric evidence. Furthermore, TAPs also provide information that cannot be obtained using traditional psychometric methods. The second research question examined accuracy of data from TAPs from the perspective of test takers. After taking part in the TAPs, students were asked to reflect on their verbalizations after they had finished their session. This information guided the investigation of verification of TAPs. The findings from this research show that students view TAPs as accurate reflections of their thought processes. Additionally, student responses provided important information about the factors that may facilitate or hinder the accuracy of TAPs. Possible implications and future directions for both research questions are discussed.

Preface

This dissertation is based on research conducted by the Assessment of Historical Thinking Project under the direction of Professors Kadriye Ercikan and Peter Seixas. On this project, I had research assistant and project management responsibilities including: preparation and collection of administrators' materials, scheduling of administrators, data collection, transcription, and data management.

The fieldwork reported in Chapter 3 was approved by the UBC Behavioural Research Ethics Board (H11-02272). The assessment tool (Appendix A) was created by K. Ercikan, P. Seixas, and Lindsay Gibson. I created the student questionnaire (Appendix B) with consultation by K. Ercikan and P. Seixas. The analyses described in Chapter 4 are my original work.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to express my gratitude to my committee members, Peter Seixas and Bruno Zumbo, for their thoughtful and insightful feedback on this dissertation. I am especially thankful to my supervisor, Kadriye Ercikan, who, in addition to her enormous support throughout this process, has been a flawless role model and mentor.

I would also like to acknowledge the ECPS administrative staff for the many ways that they have helped throughout my program.

Finally, I offer my sincere thanks to my friends in the CARME lab, as well as fellow MERM, ECPS, and Faculty of Education students. I take with me the fondest memories of our time at UBC.

Dedication

To my parents, who have always provided me with their unwavering support and love.

1 Introduction

1.1 Assessments of Complex Thinking

Engaging students in complex thinking, which builds above and beyond their basic recall or understanding of a concept, is an important educational goal. Though students should be able to recall facts that have been taught in a classroom, it is equally important that they are able to work with subject matter in a way that shows that deeper learning has also occurred. Accordingly, the assessment of complex thinking is also an important goal (Ercikan & Seixas, 2011). That is, once students have been given the skills to employ complex thinking, assessments should be able to capture those processes. Using assessments to ensure that students are engaging in complex thinking can allow educators to evaluate the learning that has taken place, provide students with feedback, motivate students to tune their complex thinking skills, appraise the quality of teaching that has occurred, sort students, and hold schools accountable for the complex thinking abilities of their students (Ennis, 1993).

1.1.1 Defining Complex Thinking

The literature surrounding complex cognitive processes, such as that which discusses critical thinking or higher order thinking, is filled with varying definitions, overlapping concepts, and terms that are sometimes used interchangeably. For instance, Bloom's taxonomy lists ascending levels of thinking, beyond the basic level of knowledge, as comprehension, application, analysis, synthesis, and evaluation (Bloom, Englehart, Furst, Hill, & Krathwohl, 1956). Likewise, one commonly accepted definition of *critical thinking* is "reasonable, reflective thinking that is focused on deciding what to believe or do" (Ennis, 1987, p.10). The integration of the elements of Bloom's taxonomy into this definition of critical thinking is clear. For

instance, the act of evaluation is implied if one is to reflect on what to believe or do. Another definition of critical thinking, which again extracts elements from Bloom's taxonomy, defines it as "the development and evaluation of arguments" (Facione, 1984, p. 259). Not only is critical thinking defined in subtly different ways, it also clearly overlaps with components of higher levels of thinking from Bloom's taxonomy.

While there is surely a need for teachers to cultivate advanced types of learning in their classrooms, the many labels given to this type of thinking can be confusing (Lewis & Smith, 1993). For the purposes of this research, because of the varying and overlapping definitions of this kind of thinking, the term *complex thinking* is used to encompass cognitive processes that require students to engage in thinking beyond simple memorization and basic understanding of information.

1.1.2 Designing Assessments of Complex Thinking

It can be challenging to design assessments that accurately capture complex thinking. For instance, past research has shown that complex thinking assessment items do not always elicit the cognitive processes that had been intended for those items (Baxter & Glaser, 1998; Koretz & Hamilton, 2006). According to Ercikan and Seixas (2011), developing an assessment of complex thinking can be challenging for two reasons. First, the authors point out that in an assessment situation, it can be difficult to design assessments that integrate both domain-specific declarative knowledge and procedural knowledge. That is, in showing understanding of a complex thinking process such as historical thinking, a student is likely required to also demonstrate some type of context-specific declarative knowledge. This relationship between declarative and procedural knowledge creates challenges in designing tasks that combine different knowledge types as well as in interpretation of student responses in relation to student development in these knowledge

types. Second, another challenge is that the assessment needs to demonstrate that complex cognitive processes are actually occurring. As was just described, test items do not always require students to engage in complex thinking, despite the intentions of test developers. This is due to challenges in determining what kinds of cognitive processes an assessment task may involve beyond simple recall (Baxter & Glaser, 1998; Ercikan & Seixas, 2011).

The challenges associated with creating assessments of complex thinking can result in assessments that do not match the goals of a curriculum. For instance, Lane (2004) reports that inconsistencies have been found between the complexity levels of state-specific content standards and statewide assessment items. In particular, she cites studies by Webb (1999; 2002) which found that assessment items were often less complex than their corresponding objective. In order to manage this mismatch of item information and objectives, it is necessary to gather suitable validity evidence that supports the inferences of an assessment.

1.1.2.1 Validity Evidence

The concept of validity is often misinterpreted as being the property of an assessment, discussed with respect to whether or not a test measures what it is supposed to measure (Borsboom, Mellenbergh, & Heerden, 2004). However, a closer inspection of the validity literature points to validity as being a property of the meaning of assessment results (Messick, 1995). Take, for example, a mathematics assessment, and the scores that are produced from students who partake in the assessment. Since the scores from a single assessment may be used in a variety of ways, such as problem solving in mathematics, test taking ability, or even teacher effectiveness, it would be problematic to attribute validity to the test itself. Rather, each inference based on test scores would require its own validation.

In a validity investigation, an argument supporting the inferences of an assessment should include several sources of evidence (Kane, 1992, 2006; Messick, 1989). *The Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) lists five sources of validity evidence: that which is based on content, response processes, internal structure, consequences, and evidence related to other variables. The evidence itself may come from a number of sources. For instance, evidence based on internal structure may include the results of a factor analysis, while evidence related to other variables may be correlations to other measures. Evidence based on content often includes expert judgment of what an assessment measures. While these individual methods are all useful for contributing to the validity of inferences of an assessment, each is limited if used on its own. Consequently, validity theorists recommend that an argument should be built using multiple sources of validity evidence. To solidify the notion that multiple sources of evidence should be used for a validity argument, take the example of using expert knowledge as evidence based on content. Though expert knowledge certainly extends beyond that of lay knowledge, and can therefore be an important component of validity investigations, it has been shown that experts may not be able to accurately predict outcomes in certain situations (Camerer & Johnson, 1991). For instance, in predicting cognitive processes used by students for an educational assessment, Gierl (1997) found that there was only a 54% match between the levels of thinking used by students for Grade 7 mathematics items, and the levels of thinking intended for those items by item developers. In this example, item developers were unable to accurately predict the cognitive processes used by students for the items, despite their presumed expertise at creating a mathematics assessment. Furthermore, Ercikan et al. (2010) found that expert judgments were not sufficient in determining bias in items that

functioned differentially. The authors conclude that additional evidence relating to student thinking processes is necessary to determine sources of bias.

1.1.3 Think Aloud Protocols

Response processes refer to the way in which an examinee understands, thinks about, and formulates answers to the items that are presented to him or her in an assessment. If a student reaches the correct answer for an assessment item, the implication is that the student was able to engage in the complex thinking that is the object of the item. Evidence based on response processes can help determine whether those processes are indeed taking place.

One way to collect evidence based on response processes is through think aloud protocols (TAPs) in which an examinee "thinks out loud" as he or she answers items (Ericsson, 2006). Rather than describing thoughts, the examinee verbalizes exactly what is going through his or her mind, to the best of his or her ability. This method is useful to researchers because it is a more direct source of data on cognitive processes compared to examinees responses to test items or experts' judgments of what kind of cognitive processes are engaged in by examinees. For example, expert judgments, which are described above as evidence related to content, may attempt to predict which cognitive processes a student uses to provide an answer to an item. However, an expert judgment is a prediction based on the item's surface characteristics and does not actually provide the response processes that a student would use in an assessment situation. As Gierl's (1997) research points out, expert judgments can sometimes be wrong about the processes in which a student will engage. A factor analysis, which provides evidence related to internal structure, uses item inter-correlations to show underlying dimensions. However, the researcher must then assign meaning to those dimensions. That is, results from an assessment may be factor analyzed to show underlying constructs based on item inter-correlations. Though a

factor analysis can show patterns that point to a number of latent dimensions, the researcher performing the factor analysis is left to determine what those dimensions represent. TAPs are potentially a way of revealing the dimensions by showing the researcher which cognitive processes are taking place among examinees. Likewise, another psychometric method of validating an assessment is to examine differential item functioning (DIF) between subgroups. DIF reveals items that function differently based on similar group ability levels. That is, given the same ability level, if one group of examinees is more likely than the other group to answer an item correctly, that item is said to exhibit DIF. A limitation to this method is that, though DIF is identified in an item, the reason for DIF may not be apparent. TAPs can compliment this method of inquiry because it can be used to identify reasons for DIF (Ercikan et al., 2010). TAPs can also be used to provide additional validity evidence when examining items of different difficulty levels. In measurement, the difficulty and discrimination parameters are properties of items that indicate how likely a student is to answer the item correctly and how well the item discriminates between high and low ability students. TAPs can add to the information about items by providing researchers with direct evidence as to why an item is easy or difficult, or even why an item may be easy for some students but difficult for others. Furthermore, items may have different difficulties for students of different ability levels. TAPs can be used to show if different cognitive processes are used for students of varying abilities.

One example of how TAPs may be used to inform and improve assessments is through evidence centered design (ECD; Ercikan, 2006; Ercikan & Seixas, 2011, Mislevy, Almond, & Lucas, 2003; Mislevy & Haertel, 2006; Mislevy, Steinberg, & Almond, 2002). This method provides a framework for designing, assembling, and delivering assessment. ECD conceptualizes aspects of the assessment process into different "models": the task model, the student model, and

the evidence model. The task model refers to the way in which a student shows what he or she knows, or the task on which a student performs. The student model is the construct that the test developer is trying to elicit. It is the latent processes in which the student engages to arrive at a correct answer. The evidence model compliments the two previously mentioned models by specifying how the student's actions and products correspond to what the student actually knows. The evidence model is comprised of two components: the evaluation and measurement models. The evaluation model refers to the rules for how an examinee's performance on a task can be assessed. The measurement model has to do with the data that is generated from student responses. In most situations, the measurement model is specified by classical test theory or item response theory (Ercikan, 2006).

Collecting evidence based on these models can help assessment developers to build a strong evidentiary argument for the purposes of an assessment. Understanding student thinking processes is one way to create a link between the construct that is being assessed and the tasks which students are asked to perform. In other words, researchers can use the TAPs to show that the student model includes aspects of cognition that are targeted by the assessment. The TAPs may also be used to demonstrate that the tasks of the assessment, or the items, are requiring students to use the cognitive processes from the student model. TAPs can provide evidence about how specific items get students to think about the construct, and highlight differences between the tasks.

1.1.3.1 Verification of TAPs

Although think aloud methods have been used for many years to examine the validity of scales and assessments, as well as issues of usability, the validity of the method itself has often been called into question. For instance, some researchers have argued that verbal reports are

incomplete measures of thought processes (Wilson, 1994). Although Ericsson and Simon (1993) do acknowledge that verbal reports are unable to capture automatic processes that occur outside the realm of consciousness, such as a reflexive response which a person does not think about in order to carry out, the authors argue that those processes are rare and not essential to the representation of thoughts that are created by using TAPs (Wilson, 1994).

One method of examining the verification of TAP data is to collect information about participants' perceptions of their own TAPs. That is, to what extent do participants feel that their verbalizations are accurate representations of their thoughts? An overview of the TAP literature shows that this question has seemingly been neglected, along with other related questions that would investigate this issue. For instance, if participants feel that their TAPs and thought processes are dissimilar, are they able to theorize possible reasons for the discrepancy? One recent study examining the usability of an online library catalogue asked participants to rate their experiences of providing verbal responses. Subjects were asked to rate the extent to which they found the experience difficult, unpleasant, unnatural, tiring, and time consuming (van den Haak, de Jong, & Schellens, 2003). However, both in the academic assessment literature and beyond, it appears that little has been done to examine how subjects felt about the degree to which their verbalizations matched their thoughts. In order to understand the relationship between students' TAPs and their perceptions of those verbalizations, the question of whether or not those perceptions are accurate is particularly relevant.

1.2 Research Questions

The research questions will examine how TAPs contribute to validating inferences from assessments of complex thinking. These research questions are:

a) How do TAPs provide validity evidence above and beyond psychometric evidence for assessments of complex thinking? That is, to what extent are the verbalizations provided by students consistent with the psychometric evidence that is collected for the assessment, and what kinds of additional information, if any, do they offer? In particular, do TAPs provide validity evidence about whether items capture complex thinking? Included in this question are two sub-questions which compare TAPs and psychometric evidence in two distinct areas:

   i. How do TAPs compare to other validity evidence with respect to the information they provide about separate gender groups? Are there differences in the quality and quantity of verbal reports for female versus male students?

   ii. How do TAPs provide information about the cognitive processes that are elicited from multiple choice (MC) versus constructed response (CR) items, and how does that information compare to psychometric evidence? Is there a difference between these types of items with respect to complex thinking, and historical thinking in particular?

b) To what extent do students believe their verbal reports are reflections of their thinking processes? Furthermore, what are sources of similarity or dissimilarity, as described by the student?

1.3 Significance of Research Questions

This work contributes to research surrounding TAPs for the validation of assessments, and particularly to assessments of complex thinking. While psychometric evidence is used frequently for validity investigations, TAPs is among one of the least used methods of validation.

The question of whether TAPs provide evidence that is consistent with psychometric evidence, and if it provides any additional information, is necessary to answer if TAPs are to be accepted as a useful validation method in the investigation of assessments of complex thinking.

Understanding students' views of their own TAPs can be extremely useful in determining whether or not verbalizations align with thought processes. As of yet, using this type of information as an indicator of precision has not been done in investigations of TAPs as an assessment validation method.

In the validity investigation for an assessment of complex thinking, researchers have a variety of procedural choices to consider. Those choices can be combined to build an evidentiary argument that supports the validity of the inferences of the assessment. This research will aid in providing researchers with insight into another important source of evidence with which to build a strong validity case.

2 Literature Review

2.1 The Relationship between TAPs and Psychometric Evidence

As discussed in the first chapter, the prevalent view of validity theory revolves around an argument-based approach (Kane, 1992; 2006; Messick, 1989). That is, evidence is collected with the intention of building a validity argument for the interpretations and uses of an assessment. In the *Standards for Educational and Psychological Testing* (AERA et al., 1999), evidence based on content, response processes, internal structure, consequences, and evidence related to other variables are the five sources of validity evidence that are described for building an argument. In educational measurement, a large proportion of validity evidence stems from evidence that is based on internal structure. This type of evidence indicates the relationship between test items and test components. In practice, this type of evidence can include methods such as factor analysis or differential item functioning (DIF; AERA et al., 1999), which is often referred to as "psychometric" evidence.

Factor analysis provides evidence at the test level by revealing dimensions, or factors, of an assessment based on item inter-correlations. This type of validity evidence is sometimes used to verify the constructs targeted by the assessment. Since factor analysis uses item interrelationships to determine factors, dimensions may sometimes be based on elements aside from the intended construct, such as the wording of a question or other characteristics of an item. DIF is a method of examining data at the item level. It identifies items which function differently for a focal and a reference group based on matched ability levels. Other psychometric evidence at the item level can include difficulty and discrimination parameters for items. Difficulty, or the "b-parameter", as the name implies, is based on the proportion of students who answer an item correctly. The discrimination parameter, or the "a-parameter", of an item indicates the degree to

which the item distinguishes between students from varying ability levels. For instance, an item with a low discrimination parameter may not adequately differentiate low ability students from high ability students, whereas an item with a high discrimination parameter would distinguish those students better. The rise of computer based statistical programs that aid in the kind of methods described here is likely the primary reason that such a large proportion of validity evidence is based on psychometric evidence.

It has been established that accumulating various types of validity evidence is necessary to adequately construct a validity argument. With that understood, it is clear that multiple types of validity evidence should point to the same message; namely, that the targeted interpretations and uses of the assessments are appropriate. Different evidence may provide different information relating to validity. As pointed out in the first chapter, some types of data such as that collected from TAPs, can provide important corroborating information for other types of validity evidence. The evidence collected in a validity investigation should overlap in a way that builds an argument for the purposes of an assessment.

To date, some studies have examined the degree to which TAPs and psychometric evidence compare with one another, with particular attention paid to the supplementary evidence that TAPs provide. For instance, Harrison, McLaughlin, and Coalter (1996) investigated item context effects of self-report questionnaires, that is, how respondent answers can be influenced by factors aside from the construct in question, using both psychometric and TAP methods. These researchers found that while psychometric methods pointed out information about mean differences, reliabilities, and relations between constructs, the TAPs provided cognitive data that explained the basis of the psychometric findings. Most notably, TAPs provided information about cognitive processes with respect to bias when responding to items.

In educational assessment contexts, TAPs have been used in combination with psychometric findings as well. For instance, Ercikan et al. (2010) used TAPs to examine sources of DIF in mathematics and science items, as determined by two separate DIF detection methods from a previous study (Ercikan, Gierl, McCreith, Puhan, & Koh, 2004). Likewise, Uiterwijk and Vallen (2005) used DIF analyses to find differentially functioning items for minority-status students on a Dutch achievement test, and then used student TAPs to investigate the sources of DIF. In another study, Arbuthnot (2009) used TAPs to understand test-taking strategies of students who showed DIF and non-DIF results in a study that investigated stereotype threat.

Despite the complimentary use of evidence in these research studies cited above, researchers have yet to use TAPs to understand psychometric findings in the context of assessments of complex thinking.

2.1.1 Response Processes as Validity Evidence

In listing response processes as one source of validity evidence, the *Standards* (AERA et al., 1999) proposes that this sort of information gathered from examinees "enriches the definition of a construct" (p.12). It also draws attention to the fact that response processes can reveal information about different subgroups of test takers, which can then lead to establishing if there are factors that affect performance differentially. The *Standards* also points out that evidence based on response processes is not limited to test takers, but rather, can be taken from observers or judges of examinee performance as well.

However, among the five sources of validity evidence listed in the *Standards*, evidence based on response processes generally receives little attention. For instance, Cizek, Rosenberg, and Koons (2008) reviewed sources of validity evidence that were used for entries in a single edition of the *Mental Measurements Yearbook*. Though a focal- and very relevant- point of the

article is that evidence based on consequences was almost non-existent in the vast review of the empirical studies, it was also found that there was scant evidence based on response processes. In fact, out of the 283 tests reviewed, only 1.8% produced evidence of response processes. Though it should be noted that this research involved the review of psychological tests rather than educational assessments, similar studies have found that response processes are underutilized in education literature as well (Shear & Zumbo, 2012). While there may be other reasons involved, the unpopularity of using response processes is likely due to the time-consuming nature of the process. For instance, Hubley and Zumbo (1996) point out that researchers may often offer convenient evidence to support the validity of a measure. Compared to many other validation methods, TAPs are anything but convenient, as they require extended time and are labour intensive.

Another factor that may hinder the popularity of TAPs is likely that it is often viewed as "soft data" (Ericsson & Simon, 1993). In other words, due to possible subjective interpretations of TAPs, it may be viewed more critically than "hard data", which is viewed as being more objective. However, Ericsson and Simon argue that verbal data can be viewed as "hard data" since technological advances in research methods allow for more precise data collection practices. For instance, instead of paraphrasing subject responses using hand written notes, including inferences that may have been used for summarization, audio and video recorders allow researchers to capture exact verbalizations that can then be analyzed in a systematic way. Furthermore, Ericsson and Simon argue that the use of information processing models as a basis for TAPs strengthens the method, making it less ambiguous to researchers. A description of the mechanisms of information processing is provided later on in this chapter.

Gaining insight into how TAPs are able to provide unique and robust evidence for assessments may help to break down these barriers in using the method for validation purposes. Increasing the use of TAPs in validation studies is one step toward an explanation-focused approach to validity research, which has been advocated for by leading validity researchers (Messick, 1989; Zumbo, 2009).

2.2 Complex Thinking

The ordering of thinking skills has enjoyed a great deal of discussion in educational research in the past century. One of the most recognized hierarchies of cognitive skills is Bloom's taxonomy (Bloom et al., 1956), which lists knowledge, comprehension, application, analysis, synthesis, and evaluation as thinking skills in ascending order. The ability to analyze and apply information to new concepts, compare and contrast, combine information from various sources, and determine quality or worth are thought to be *higher order thinking* (Airasian, Englemann, & Gallagher, 2007). *Critical thinking* is a term that is sometimes used interchangeably with higher order thinking (Giancarlo-Gittens, 2009; Leighton, 2011a). Within the critical thinking community, one commonly accepted definition is that of Ennis (1987), who describes critical thinking as "reasonable, reflective thinking that is focused on deciding what to believe or do" (p.10). Paul (1992) defines critical thinking as "disciplined, self-directed thinking that exemplifies the perfections of thinking appropriate to a particular mode or domain of thought" (p.9). However, Lewis and Smith (1993) make a point to separate the two terms, suggesting that the definition of higher order thinking must include components which are traditionally left out of critical thinking, notably creative thinking as well as problem solving.

In disciplinary areas, such as history or science, researchers argue that different types of higher order thinking is expected (Miri, David, & Uri, 2007; Peck & Seixas, 2008). *Historical*

thinking refers to higher order thinking in history and includes the ability to establish historical significance, use primary source evidence, identify continuity and change, analyze cause and consequence, take historical perspectives, and understand the ethical dimension of historical interpretations (Peck & Seixas, 2008; Seixas, 2009, as cited in Ercikan, Seixas, Lyons-Thomas, & Gibson, 2012, p. 3).

Given the various definitions provided for terms here, and the obvious overlap, it is no wonder that Cuban (1984) referred to this as a "conceptual swamp" (p.676). As established in the introductory chapter, *complex thinking* is used to encompass these terms listed above, which exclude cognitive skills that simply require a student to recall in a testing situation.

2.2.1 Assessment of Complex Thinking

As the need to teach complex thinking grows in schools, the need to assess this kind of thinking is necessary as well. Knight (1992) posits that tests of complex thinking should include items which present students with "fresh problems to solve, hypothetical situations in which to describe behavior, documents from which to draw conclusions, [and] arguments to analyze that contain errors of fact and/or reasoning" (p.70).

Including items that invite students to utilize complex thinking may be one thing, but ensuring that those questions are actually eliciting those types of thinking skills is another. Leighton (2011a) highlights that, concerning higher order thinking, tests need to be appraised for their capacity to evoke expected cognitive skills because some assessments may not always elicit from students the type of thinking that is expected. While this may be necessary, methods for evaluating the assessments may vary. For example, Cromwell (1992) suggests including expert judgments by, for instance, experienced teachers in assessments of complex thinking. However, another way of ensuring that complex thinking is taking place is to know, on good authority, the

thinking process of the student. Utilizing TAPs would therefore be one sensible method of making certain that particular cognitive processes are taking place. The benefit of using TAPs in a validity investigation is that it provides a type of evidence that is explanatory in nature (Messick, 1995). That is, in the context of an assessment of complex thinking, TAPs provide insight into the thinking processes which explain a student's answer. Therefore, TAPs constitute an explanation-focused (Zumbo, 2009) method of obtaining validity evidence for assessments of complex thinking.

2.2.1.1 ECD for Assessments of Complex Thinking

ECD has become a promising method of designing assessments of complex thinking. This approach to assessment design requires explicit identification of intended inferences and uses of assessment results. In assessments of complex thinking the link between assessment tasks, the targeted complex thinking and intended interpretation need to be made explicit. Developed by researchers at Educational Testing Service (Mislevy, Steinberg, & Almond, 2002), this method consists of three models. The student model of ECD corresponds to the construct that the assessment is meant to measure, and the task model is the way in which the assessment is designed to measure the construct. The evidence model determines how the observations of student performances are linked to the inferences that are made about what students know (Ercikan, 2006; Ercikan & Seixas, 2011). The evidence model is comprised of the interpretation model, which specifies how student responses are evaluated with respect to the construct, and the measurement model, which defines how student performance across tasks are summarized (Ercikan, 2006).

Clear interrelationships between the three models are necessary for making valid inferences about the assessment results (Ercikan & Seixas, 2011). Therefore, TAPs can be

particularly relevant to assessments developed based on ECD because they are a way of providing direct evidence about the interconnections between the construct that is intended to be measured (the student model) and the tasks that are included in the assessment (the task model).

2.3 A Brief History of TAPs

In their book, *The Think Aloud Method: A Practical Guide to Modeling Cognitive Processes*, van Someren, Barnard, and Sandberg (1994) report that the TAP method is rooted in early twentieth century psychological research. Specifically, the authors point to introspection as the original practice through which the approach was formed. The introspection method allowed subjects to orally reflect on their own thoughts and sensations. However, this method of investigation largely fell out of favor in the field of psychology, mainly because it lacked the empirical nature of other methods, such as behaviorism, which can be observed, interpreted, and generalized. As the less empirical practice of introspection was replaced by behaviorism, so too did the practice of understanding cognitive processes by subject verbalizations (van Someren et al., 1994). During this time, verbal reports were the target of criticisms that are not dissimilar to arguments against the methodological validity of TAPs today. Verbal reports were thought to be incomplete and not consistent with other observable behaviors, as well as distracting to the subject, resulting in a change of the cognitive process (Ericsson & Simon, 1993). However despite these practical criticisms, verbal reports did not completely disappear. Rather, the methodology of verbalizations was gradually modified in a way that would provide more accurate data. As think aloud methodologies continue to be refined to this day, it is argued that the technique is experiencing a resurgence across various areas of study, and academic assessment in particular.

2.4 Variations of TAPs

2.4.1 Alternate Terms and Methods Related to TAPs

The process of collecting verbalizations that reflect a subject's thought processes can go by a number of different terms and is used in a multitude of ways (Nielsen, Clemmensen, & Yssing, 2002). TAP is the designation that will be used in this paper, however *verbal reports* and *verbal protocols* are also used by researchers to describe the same procedure (Ericsson & Simon, 1985). Another technique that is related to TAPs, but dissimilar at its core, is *talk aloud protocols*. This refers to the process of having subjects verbalize their actions, but not their thoughts (Ericsson & Simon, 1993). In some cases, *think aloud* and *talk aloud* are used interchangeably because researchers choose not to distinguish between a subject's actions and his or her thoughts. However, in the case of validity research in academic assessment, the distinction is necessary. Another term that is sometimes used interchangeably with TAPs is *cognitive laboratories (labs)*. For example, Johnstone, Bottsford-Miller, and Thompson (2006) appear to use the term *think aloud method* interchangeably with *cognitive laboratories* in their paper that examines how the procedure can be used to improve large-scale assessments. In their technical report on the procedure, Zucker, Sassman, and Case (2004) portray cognitive labs as being akin to verbal reports described by Ericsson and Simon (1993). The authors also indicate that verbal reports, taken both during and after the subject completes a task, can be combined with other behavioral data from the session, such as the subjects' use of a pencil and paper while completing a test item. Additionally, the term *cognitive interview* is used by Desimone and Le Floch (2004) to include the think aloud method, which may also include accompanying retrospective probes. The authors argue that cognitive interviews are an important component in

the development of surveys in educational research, and should be utilized to better understand complex thought processes.

Two general methods of collecting verbal data exist under the overarching umbrella of TAPs. One method, which involves collecting data while subjects are working through a problem, is referred to as *concurrent TAPs*. The other method involves having subjects recount their thought processes after attempting a problem, and is known as *retrospective TAPs.*

2.4.2 Concurrent TAPs

In order for a researcher to conduct concurrent TAPs, the researcher instructs the subject to "think aloud", and therefore verbalize his/her thoughts immediately as they come into the subject's awareness. In the preface to the revised edition of their influential book, *Protocol Analysis*, Ericsson and Simon (1993) use the example of how a subject may concurrently think aloud while solving a multiplication problem. The authors explain that "a subject given the task of mentally multiplying 24 by 36….might verbalize: "36 times 24," "4 times 6," "24," "4," "carry the 2," "12," "14," "144," and so on" (p.xiii). Generally, the researcher would not prompt the subject in any way, other than to "remember to think aloud" when he or she falls silent. This act of remaining reserved on the researcher's part is carried out so as not to cause any interference during the TAP (Ericsson & Simon, 1993; van Someren et al., 1994). That is, the researcher is mindful that anything he or she says could have an influence on the subject's verbalizations, performance, or both.

2.4.3 Retrospective TAPs

In contrast to concurrent TAPs, retrospective TAPs requires the subject to *recall afterward* what (s)he had been thinking when (s)he was completing the task. The subject then verbally relays that information to the researcher. Ericsson and Simon (1993) argue that if the

retrospective TAP is performed within a reasonable amount of time to a task of short duration, the subject should be able to accurately recall his or her thought process in the true sequence that it originally occurred. For this reason, the retrospective TAP could be preferable to the concurrent TAP in some situations because it would remove the drawback of having a subject keep up with thoughts that are too quick to be verbalized. Furthermore, the authors also note that retrospective TAPs can be ideal in situations where concurrent TAPs are impractical or even dangerous, such as perceptual motor tasks.

2.4.3.1 Stimulated Recall

A technique that falls under the category of retrospective TAPs, but varies slightly from simply asking the participant to remember and verbalize his or her thoughts, is *stimulated recall* (SR). This procedure is done with the aid of a recording device that prompts the subject to recall his or her thoughts at a particular moment while viewing a video of the task being performed (Lyle, 2003). One obvious concern with this method is that the subject observing the video will "fill in" partial recollections of the event, or even react to, rather than recall thoughts from, the recorded images (Lee, Landin, & Carter, 1992; Tjeerdsma, 1997). Despite these concerns, SR is a common research technique used in teaching (Calderhead, 1981; Meade & Mcmeniman, 1992), counseling (Martin, Martin, Meyer, & Slemon, 1986), sports (Lee et al., 1992; Lyle, 2003), and nursing (Liimatainen, Poskiparta, Karhila, & Sjogren, 2001). The prime benefit of this technique is that it allows subjects to perform in a naturalistic setting and then recall thought processes after a task has been completed (Lyle, 2003).

2.4.3.2 Verbal Probing

Another method of collecting retrospective TAPs is to use a technique known as *verbal probing* (Willis, DeMaio, & Harris-Kojetin, 1999). This method involves directly probing a

subject after (s)he has finished completing a task. That is, the questions that a researcher may ask goes beyond simply asking the subject to recall her thoughts, but rather inquires about specific parts of the problem solving process, including asking the subject to paraphrase questions, clarify understanding of particular parts of the question, and report the extent to which they are confident in their answers (Willis et al., 1999).

2.4.4 Descriptions of Thinking

An important distinction to point out is the difference between methods of *thinking aloud* and *explanations of thinking*. While the concept of thinking aloud itself contains a certain degree of ambiguity (Neilson et al., 2002), it is generally accepted that verbalizing one's thoughts as they occur is separate from describing and explaining the thoughts that occur. It is a common concern that asking a subject to verbalize his or her thoughts while completing a task runs the risk of altering that person's line of thinking (Wilson, 1994). While Ericsson and Simon (1998) vehemently argue that there are "circumstances where verbalization…can be made without reactive effects" (p.178), the soundness of this concern will be discussed at greater depth further on in this chapter.

2.4.5 Comparisons between Concurrent and Retrospective TAPs

Though Ericsson and Simon (1993) recommend that when possible, both concurrent and retrospective TAPs be collected, this may not always be feasible in a realistic research setting. The researcher's decision to use either retrospective or concurrent TAPs can have significant consequences on the results of the study because of the nature of each technique. For instance, some researchers have raised concerns that the cognitive process itself changes when a subject is asked to concurrently verbalize his thoughts (Wilson, 1994). However, by using retrospective TAPs, the researcher risks having the subject forget, altogether or partially, the thought processes

that were involved in a problem-solving task (Teague, De Jesus, & Nunes-Ueno, 2001 as cited in van den Haak et al., 2003). Additionally, another concern is that subjects will reorder their thoughts when verbalizing retrospectively, mixing the past with the present (Kuusela & Paul, 2000). For these reasons and others, it is possible that the kind of information elicited from concurrent and retrospective TAPs may be different, and could consequently lead a researcher to different conclusions based on the technique that is employed. The researcher must also decide on the practicality of employing one TAP method over another, or using both. One major concern about using TAPs is the great time commitment that is required from participants who verbalize their thoughts. The duration of a retrospective TAP data collection session can often be longer than a concurrent TAP session, because the participant is first asked to perform a task and then asked to consider her cognitive process in retrospect (van den Haak, 2003). Asking subjects to concurrently verbalize thinking for items that require complex thinking, as is proposed in this research, may increase the risk of mentally wearing out participants, which in turn could result in lower quality verbalizations. Alternatively, asking students to verbalize their thoughts retrospectively risks having students forget components of the complex thinking that had occurred. Furthermore, retrospective TAPs that are aided, for instance by a video recording, may be less effective in many assessment situations compared to a physical performance of a task.

Given these various benefits and drawbacks of using concurrent versus retrospective TAPs, one important area of research from the literature is comparing the two methods of collecting TAPs. That is, some researchers have examined the differences in verbalizations when concurrent TAPs versus retrospective TAPs are used. For instance, van den Haak et al. (2003) found that in usability research, concurrent and retrospective TAPs revealed similar problems, despite those issues being revealed in different ways. That is, concurrent TAPs *demonstrated*

usability troubles, whereas retrospective TAPs revealed the same problems by means of the verbalizations. In another paper that directly compared the two methods of TAPs, Kuusela and Paul (2000) determined that concurrent TAPs generally outperformed retrospective TAPs in that they provided a greater number of verbalizations, as well as more information about the decision making process. From the perspective of a researcher investigating complex thinking, that would likely place concurrent TAPs in a more valued position over retrospective TAPs. Kuusela and Paul (2000) also determined that retrospective TAPs have the advantage of providing more statements about final choice. Again, in the case of a researcher who is interested in investigating complex thinking, understanding the process of thinking is likely more valued than the final answer, in which case concurrent verbalizations would be more appropriate if one method were chosen over the other.

2.5 Elements of TAP Methodology

2.5.1 Item Characteristics

An important consideration of TAP methodology is the type of item that a subject is expected to work through and verbalize. In the area of academic assessment, items can vary in terms of difficulty, subject domain, and type, in addition to whether or not they require students to employ complex thinking skills. These differences may have dramatic impacts on the types of verbalizations that are provided when a research participant works through a task.

2.5.1.1 Item Difficulty

It has been found that the difficulty of an item can affect aspects of verbalizations that are provided for that item. For instance, Leighton (2013, 2011b) found that item difficulty had an effect on students' nervous speech when providing TAPs while completing math problems.

Specifically, easier items were more likely to elicit irregularities of speech flow, and easy and moderate items were associated with more validation seeking comments. The author also found that easy and difficult items, as opposed to moderate items, were associated with inconsistencies in the cognitive model scores that were determined by the authors. In addition to these findings, van den Haak et al. (2003) report that tasks of moderate difficulty are often ideal, however the authors speak to task difficulty in the context of usability testing. Nonetheless, their findings also point to the relationship between task difficulty and TAPs. The authors report that using retrospective TAPs rather than concurrent TAPs may be helpful in overcoming issues such as reactivity and completeness of verbalizations that are associated with high task difficulty. In any case, the implication that moderately difficult items are ideal for TAPs is consistent with Ericsson and Simon (1993). Van Someren et al. (1994) suggest utilizing the knowledge of a domain expert to ensure that the items used for TAPs are appropriately difficult for participants.

It is important to note that items that require students to partake in complex thinking are not necessarily categorized as difficult. Rather, complex thinking items should be considered to include a range of difficulty depending on the individual item. Difficulty of an item, including those that require complex thinking, can be indicated by information such as p-values (proportion of students who answer the item correctly) or expert judgments.

2.5.1.2 Item Domain

In addition to item difficulty, a related concern that a researcher must consider is the subject domain that a TAP item assesses. A researcher validating the interpretations of a mathematics exam would logically provide mathematics items for TAPs. In a 1985 chapter, Ericsson and Simon point out differences that exist between having participants verbalize their thoughts when working through multiplication problems, which involve numeracy skills, versus

anagrams, which require literacy skills as it involves providing subjects with scrambled letters and asking them to form a word from those letters. In particular, the authors show that the kinds of verbalizations collected from the two types of problems are very different. When given a multiplication type problem, the respondents usually addressed the task using the same sequence of operations and results. However, when given anagram-type problems, respondents used different information or different ordering of steps. Respondents used recognition and often evoked information from long-term memory, so the verbalizations had more variation. Though the findings from this research should not be generalized to all mathematics- or linguistic-based TAPs, it does point out that the type of task presented to a subject can influence the kind of verbalizations that may result.

2.5.1.3 Item Type

The purpose of TAPs is to ensure that items are requiring students to use the type of thinking that they intend to measure. In order to do that, the researcher must be clear on what type of thinking he or she is interested in investigating. That is, the goal of an assessment may range from lower level thinking to complex forms of thinking. MC type items are more likely to measure recall (Airasian et al., 2007), however in some cases, they can also assess complex thinking. For instance, Norris (1989) claims that, while MC tests cannot measure all aspects of complex thinking, carefully thought out MC items may be able to measure certain facets of complex thinking, namely judgments of credibility.

Beyond levels of thinking, some have argued that using TAPs specifically for different item formats is important. Gorin (2006) points out that verbal responses should be used in the validation of assessments, and that examining responses to different item types, such as MC and CR, can allow test developers to better understand domain-level processes rather than processes

that may be item specific. Thus, understanding how students respond to different item types would be imperative. That is, if students provide fundamentally different types of verbalizations to one type of item over another, caution would have to be given to the interpretations of those protocols. In addition to the specific type of item that is used for TAPs, the presentation of items may be important. Harrison et al. (1996) report that when subjects were asked to verbalize their thoughts when answering items of an organizational justice questionnaire, respondents rushed through their answers when the items were presented in a way that filled the entire page. When items were spread out with more white space per page, participants were more likely to allot extra time and deliberation to their verbalizations.

Given that there is little research literature that examines the effect that item type has on TAPs, this will be investigated in the present research, as outlined in the research questions introduced in the first chapter.

2.5.2 Interviewer Effects

Another aspect of TAP methodology that can have an effect on performance, as well as the types of verbalizations produced, is the effect of the interviewer. That is, certain characteristics of the researcher can influence the way in which the subject approaches questions and verbalizes those answers. In their discussion of verbal protocols, Ericsson and Simon (1993) suggest that the interviewer not be present while the subject produces his or her verbalizations (as cited in Leighton, 2013). However, not being present is simply not realistic in many research settings. Alternatively, van Someren et al. (1994) suggest that the interviewer remain reserved throughout the interview in order to avoid interference. In other words, other than the prompt to "keep thinking aloud" when the subject falls silent, interviewers should restrain themselves from speaking at all other times.

2.5.2.1 Expertise of Interviewer

Some research indicates that there may be qualities of the researcher, separate from his or her interactions during the session, which affect the TAPs of the subject. For instance, Norris (1990) reports that interviewer effects were found to affect student performance scores in a TAP study involving MC assessment, despite efforts to standardize interview procedures. Unfortunately in that particular paper, specific characteristics of the interviewers were not reported. Other research has pointed to the influence that interviewer knowledge has on TAP subjects. For instance, when investigating student TAPs collected during a math achievement test, Leighton (2013, 2011b) found that there was a significant difference in performance when the interviewers identified themselves as experts in the subject domain, as well as when students assumed the interviewer was an expert. However, with regard to their verbalizations, students did not exhibit nervous speech across expert/non-expert interviewer conditions. Nonetheless, the author hypothesizes that the presence of an expert interviewer caused students to feel a certain degree of anxiety, which influenced item performance.

2.5.2.2 Gender of Interviewer

Another interviewer characteristic that has received attention in TAP methodology research is the gender of the interviewer. Leighton (2013, 2011b) investigated the effect of male versus female interviewer effects on both student item performance and TAPs. However, findings from the study did not show an effect of interviewer gender on item performance or nervousness.

2.5.3 Ideal Subjects

In addition to the nature of the tasks and possible interviewer effects, another important methodological aspect to consider is the characteristics of the think aloud participants.

Consideration of subject attributes may be important as those characteristics may have an effect on the verbalizations that participants produce. For instance, Harrison et al. (1996) report large deviation on the amount of vocalizations that were provided by subjects. However, the authors failed to investigate possible sources of that variation. Van Someren et al. (1994) point out that subjects taking part in a think aloud activity ought to be a random sample from the population of people that the research is meant to address. In the case of academic assessments, a school-age population is often the target. Thus, TAP participants should be a representative sample, and generalizations may not be appropriate for people outside of this population. For instance, if TAPs were being used for validation of a high school exit exam, then using participants who are of the appropriate grade and who have completed relevant coursework would be the only suitable subjects to include. If for any reason the sample is not representative of the population, the researcher should be aware of the limitations of the research.

2.5.3.1 Sample Size

One unresolved issue in the area of TAPs, and qualitative research in general, is the sample size needed for adequate data collection. In the "sample size" section of their article describing TAPs, Fonteyn, Kuipers, and Grobe (1993) report that the think aloud method "seeks rich, in depth data from a small sample" (p. 432) without providing a range or minimum number. Nielsen (1994) recommends in his paper, *Estimating the number of subjects needed for a thinking aloud test*, that 4±1 subjects are required for a TAP study. In that article, the author describes his study in which subjects are given tasks to complete, and the resulting recommended sample size is reflective of how many people it would take to find ~75% of the usability problems. However, as discussed earlier, TAPs for usability research are not necessarily generalizable to TAPs for the purposes of validation of academic assessments. While some

findings may translate to educational assessment research, other findings may not. In this case, determining "usability problems" is one issue, but understanding alternative thought processes may be a research goal, especially depending on the type of thinking that is the objective of the research.

2.5.3.2 Linguistic Ability of Subject

One possible research goal in TAP research is to understand the translation or language issues associated with an assessment (Ercikan et al., 2010; Uiterwijk & Vallen, 2005). For instance, researchers may want to verify that a test has been translated in a way that the construct being measured is equivalent in both language versions. Especially significant to this type of research, van Someren et al. (1994) list verbalization skills as a relevant issue when determining sample participants. Thus, language ability of subjects should be one factor to consider in a multicultural or multi-linguistic study. This issue can arise in multiple ways. For instance, if bilingual subjects are used to elicit verbalizations, a researcher would want to ensure that the subjects can accurately express their intentions in the language of the test and TAP sessions.

In order for TAP research to be generalizable to a larger group, the TAP sample should be representative of the population of interest. For example, in their research examining the sources of DIF in French and English versions of a Canadian national assessment, Ercikan et al. (2010) acknowledge that their French-speaking sample was limited in that the students lived in English-speaking environments, thus resulting in possible "poorer French language competencies than the DIF sample" (p.34). In this case, the DIF sample included French-speaking students living in majority French-speaking environments. However, the authors point out that the DIF sample also included French-speaking students from an English-speaking environment, so the

alternative of using TAPs solely from French-speaking students who live in a French-speaking environment would have been inappropriate.

2.5.3.3 Age of Subject

Another issue to consider when carrying out TAPs, related to verbalization skills, is the age of participants. In academic assessment, a wide range of school-age students may be targeted for validation studies. However, the age of students may have an effect on aspects of their articulations. Charness (1981) used TAPs from subjects between ages 16 to 64 to understand thought processes of chess positions, paying attention particularly to participant differences in age and skill. The findings suggest that age was sometimes a factor in chess strategies; however, the authors did not indicate that variations in age had any effect on subject verbalizations.

In terms of the ability of children to verbalize thought processes, van Someren et al. (1994) report that "[y]oung children usually find it difficult to think aloud" (p.36). However, the specific age at which children move beyond difficulties verbalizing their thoughts is unclear. Past research has used, and even compared, TAP data from primary school-aged children. Cremeens, Eiser, and Blades (2007) used TAPs to understand children's thinking when answering a quality of life measure. The authors report that TAPs revealed strategy differences as a function of age, with older children (7- to 9-year-olds) more likely to describe social comparisons and use concrete examples, and younger children (5- to 6-year-olds) not giving reasons for their responses. Likewise, Gaderman, Guhn, and Zumbo (2011) also used TAPs with children to understand thinking about quality of life items, and report that age differences were associated with different response strategies. Again, it should be noted that this research was not investigating TAPs methodology with respect to age, but rather used TAPs to understand children's thinking about quality of life.

2.5.3.4 Expertise of Subject

Another subject characteristic that the researcher must consider is the extent to which the participant is knowledgeable about the task (s)he is being asked to complete. In some situations, insight into an expert thought process is ideal. For instance, Ercikan et al. (2010) point out that TAPs could be used for "understanding differences in novice versus expert performances in a variety of areas, such as chess, music, physics, sports, and medicine" (p.25). Likewise, Ericsson (2006) points out that expert behaviors may sometimes appear counterintuitive to others. The author indicates that verbalizations are an ideal methodology to understand the motivations of experts, as well as mediating factors of expert performance. In the validation of academic assessments, "expert" versus "novice" could potentially be translated into high achieving versus low achieving students. Student familiarity and problem solving abilities should be taken into account when TAPs are collected. That is, if students have characteristics that may indicate advanced knowledge of a subject, consideration should be given to the possibility that those students may think about problems differently than lower achieving students. Leighton (2013, 2011b) found that prior math achievement had an effect not only on item accuracy, but on the level of cognitive models that students exhibited as well.

2.5.3.5 Gender of Subject

A final participant variable that may affect TAPs is gender of the subject. There has not been a great deal of literature up to this point that specifically investigates the differences of TAPs between males and females, and the literature that is available points to conflicting evidence. For instance, Klinger (1971) noted that only female subjects who were deemed to be highly verbal were able to provide articulate verbalizations in a visual images task (as cited in Ericsson & Simon, 1993). However, in another study, Norris (1990) found that there were gender

interaction effects in favour of males when high school students were asked to provide verbalizations for a critical thinking test. Specifically, the author found that male students had significantly higher mean thinking scores when providing TAPs for one interviewer but not another. However, Norris does not provide a reason for this effect. In other research, Leighton (2013, 2011b) initially sought to explore the role that students' gender had on TAPs, though preliminary analyses showed that it did not have an effect on accuracy or interact with other variables being investigated in the research. For that reason, student gender was omitted from further analyses.

One consideration related to gender differences and TAPs is whether there is evidence that one gender is more verbal than another, and would therefore be more likely to produce more verbalizations. Influential work by Maccoby and Jacklin (1974) suggest a consistent female advantage in verbal ability. In a later meta-analysis on the same topic, Hyde and Linn (1988) also found a slight female advantage in verbal ability. However, the difference was so small that the authors dismiss the idea that a gender difference persists, if it had ever been present.

Differences between the ways in which females and males *verbalize* their thoughts versus differences in *thought content* (that is subsequently verbalized) can be difficult to tease apart. Some research has revealed differences in how genders verbalize their thoughts, despite similarities in the actions that participants take while verbalizing. In one study that asked participants to think aloud as they found their way back to a location using an unfamiliar route, the authors found that females reported being more uncertain about their route, despite the fact that there were not gender differences in the choice of route (Lawton, Charleston, & Zieles, 1996). However, again, it is not possible to disentangle if the gender differences were more strongly associated with way-finding or the act of verbalizing one's thoughts.

2.5.4 Setting Up a Session

Once certain aspects of the TAP session are acknowledged and dealt with, such as item types, participant characteristics, and interviewer effects, the researcher can then proceed with running a TAP session. Though the methods used to obtain TAPs can differ considerably (Kuusela & Paul, 2000), Payne (1994) lists some "best practices" that he believes best facilitate think aloud verbalizations from respondents. Among them, Payne suggests introducing the session with a practice item so that subjects are able to fully understand the task beforehand. Doing so, the author suggests, is one way to ensure that the subjects are comfortable with thinking aloud while performing a task and ensuring that the thinking aloud is secondary to the actual performance. Ericsson and Simon (1993) also advocate for a warm up task at the start of a session, pointing out that this practice can train the subject to conform to TAP directions. Payne (1994) suggests that another "best practice" would involve encouraging the subject to verbalize all thoughts that occur while performing a task and not prompting the subject to expand on specific information. Likewise, Fonteyn et al. (1993) suggest that the only utterance a researcher should make during a session is a reminder to "keep thinking aloud". Van Someren et al. (1994) echo this sentiment, indicating that the experimenter should only remind the TAP participant to "keep on talking", despite the potential urge to become more involved, and Leighton (2004) suggests that when prompts are necessary, they should be neutral and infrequent, should only be offered after 10 to 15 seconds of silence have passed, and researchers should avoid interrupting students during their thought processes. Finally, Payne (1994) suggests that during the session, the researcher should not be visible to the participant so as to avoid any interactions between the two people. While Ericsson and Simon (1993) also recognize that this may be the route for some researchers, the authors point out that the researcher may need to be present in order to prompt a

subject if (s)he falls silent while providing verbalizations. Again, using a variation of the sole prompt, "keep thinking aloud", would also lend itself to ensuring that interferences from the researcher to the subject is minimized

2.6 What is Captured in TAPs?

Given the great deal of attention given to various aspects of TAP sessions, one glaring question arises: What exactly is being captured? That is, when subjects are asked to verbalize their thoughts, how close are those verbalizations to cognitive processes, and what sort of theory supports the claim that TAPs may be used as evidence of thinking?

2.6.1 The Information Processing Model

Ericsson and Simon (1993) argue that the information processing model supports why TAPs are a useful method of understanding some cognitive processes. The primary assumption of the information processing model, and why TAPs can be a window into cognitive processes, is that all knowledge of which we are conscious has to go through our short-term memory (STM). Once that information has passed into the STM, and before it leaves the STM, we are able to verbalize it. TAPs are the verbalizations of information that is stored in the STM. However, if too much time has elapsed, and the information has left the STM, the resulting verbalizations will be descriptions and explanations of cognitive processes, rather than TAPs (Nielson et al., 2002). That is to say, *thinking aloud* has a limited time frame in which it can be captured; its presence in the STM determines the margin of time.

2.6.2 Validity Issues

It was briefly mentioned earlier in this chapter that some researchers have expressed concern that concurrent verbalizations of a subject's thoughts during a task may not be fully

reflective of the natural cognitive processes of that subject. With respect to retrospective TAPs, a common concern is that verbalizations would be "sanitized" versions of thoughts rather than the actual thoughts themselves. In either case, the information suggested from the TAPs would not necessarily be accurate because the verbalizations would not be precise reflections of naturally occurring thought processes. In other words, the validity of the methodology itself would be compromised. Additionally, Payne (1994) points out that the validity of the protocols can depend on the tasks from which they are collected.

A major concern is the "completeness" of a respondent's verbalizations (Wilson, 1994). For instance, concurrent verbalizations may be somewhat incomplete if a respondent is providing TAPs for a task which involves thinking that is peripheral to the subject's attention or cannot be readily articulated (Ericsson and Simon, 1993, as cited in Wilson, 1994). Van Someren et al. (1994) adds that verbalizations may be incomplete because the speed at which thoughts occur may be faster than the time it takes to report them verbally.

2.6.2.1 Reactivity

A major point of contention surrounding concurrent TAPs is the possibility that simultaneous verbalizations can change the course of thought processes. For instance, Russo, Johnson, and Stephens (1989) list reactivity as a major source of TAP invalidity. The authors list four possible reasons for reactivity: "(1) the additional demand for processing resources, (2) auditory feedback, (3) enhanced learning over repeated trials, and (4) a motivational shift toward greater accuracy" (p. 764). In an investigation that involved having subjects answer verbal, numerical, pictorial, and mental addition tasks, the authors report that not only do concurrent verbalizations have a negative effect on task accuracy, but they also result in prolonged response times.

However, some researchers are also of the opinion that reactivity is not a legitimate source of invalidity. Leow and Morgan-Short (2004) investigated the potential reactivity of concurrent TAPs by having second language acquisition learners perform tasks in think aloud and control groups. The authors report that concurrent TAPs did not have detrimental or facilitative effects on subjects' comprehension, intake, or written production. Additionally, when examining an attention task related to sport, Williams and Davids (1997) maintain that asking participants to verbalize their thoughts did not have an effect on task performance. In another study, Henry, LeBreck, and Holzemer (1989) investigated the effect that verbalizations had on performance of a computerized clinical simulation. Participants were divided into three groups: one group was instructed to concurrently think aloud, another was asked to recollect decision processes after the task was completed, and the third group did not verbalize any processes. It was found that the groups did not differ on either efficiency or proficiency, leading the authors to conclude that concurrent verbalization does not have an effect on cognitive processes. Furthermore, analyses of the data indicated that experience level was irrelevant to the findings. In another study that used TAPs to investigate participants' cognitive processes while completing an organizational justice questionnaire, the authors concluded that survey responses yielded no differences in means, variances, or covariance patterns between TAP and non-TAP subjects. The authors conclude that asking subjects to verbalize their thoughts had no effect on the cognitive processes of those subjects (Harrison et al., 1996).

These reports are directly in line with Ericsson and Simon's (1985) claim that no differences in performance exist when employing think aloud verbalizations. The only exception to that claim is their own finding that performance speed changes for subjects whose thinking involved non-verbal codes.

2.6.2.2 Are Retrospective Verbalizations "Sanitized"?

As an alternative to concurrent TAPs, retrospective TAPs provide the benefit of allowing participants to finish the task before verbalizing their thoughts. As a result, synchronization of thoughts and protocols is not an issue, nor is the immediate possibility of the thought process being disturbed. However, it is feasible that the TAPs generated from this method are not fully reflective of mental processes. Ericsson and Simon (1993) point out that retrospective TAPs, compared to concurrent TAPs, are more vulnerable to modifications of the actual thought process. Russo et al. (1989) refer to this as being *nonveridical,* which is caused by either leaving information out or supplementing cognitive processes that did not take place. In their paper, Russo et al. report that retrospective TAPs from their investigation demonstrated forgetting and fabrication of cognitive processes. As well, Leighton (2013, 2011b) reports that her findings suggest that retrospective TAPs measure *idealized* problem solving. That is, participants in her study were more likely to report problem solving that was thought to be viewed as more impressive to the researcher than actual problem solving methods.

Another point to consider about verbal reports is introduced by Nisbitt and Wilson (1977), who argue that researchers may not have access to some types of cognitive processes. For instance, the authors point out that when collecting verbal reports, there may be stimuli that are not perceived by subjects, despite the influence of those stimuli on responses. Instead, the authors maintain that subjects may base their own reports on pre-existing beliefs about the cause and effect of their responses. Consequently, it is argued that true reports of cognitive processes can only occur when stimuli are fully noticeable to the subject. Ericsson and Simon (1993, as cited in Wilson, 1994) do not disagree that some TAPs will, at times, be incomplete. The authors

acknowledge that TAPs will not be able to provide information about (1) processes that are automated and (2) non-verbal thoughts that cannot be converted into verbal code.

2.6.3 The Relationship between Language and Thought

An important aspect to consider when using TAPs for validation purposes is the relationship between language and thought. For instance, Ericsson and Simon (1998) argue that when thoughts are in oral form, no additional processes are needed to produce verbalizations. However, when information is in non-oral form, the authors agree that additional processes may be needed, which would increase the time between the thought occurring and a verbalization. However, it is also argued that the cognitive process itself does not change with respect to task performance. Conversely, in a study that examined the effect of verbalizations on problem solving, Schooler, Ohlsson, and Brooks (1993) argue that having participants think aloud when completing non-reportable tasks results in decreased performance.

2.6.4 Participant Perceptions of TAPs

One essential concern that has been briefly touched on in the literature is how the participants themselves perceive TAPs. That is, while researchers may attempt to determine what is being measured by TAPs, and the validity of the method itself, participants may also have keen insight into the verbalizations which they produce. Currently, few research studies appear to have consulted their research subjects about their opinions of what they reported to have been thinking. Harrison et al. (1996) report that subjects in their TAP study did not express "concern, discomfort, or difficulty" (p.253) with providing think alouds, however do not go into detail to explain if participant reactions were elicited by the researchers. In one study that did ask for participant feedback on TAPs, van den Haak et al. (2003) asked participants to rate their experiences of both concurrent and retrospective TAPs, as well as participants' ease of thinking

aloud (i.e., the extent to which it was difficult, unpleasant, tiring, unnatural, and time consuming). The participants rated their experiences rather neutrally, with no significant differences between the concurrent and retrospective TAP groups. On a one to five point scale with one being positive and five being negative, the participants indicated the difficulty/ease of performing TAPs as 2.4 and 2.7, for concurrent and retrospective TAPs respectively. It was reported that unpleasantness/pleasantness of performing TAPs was 2.7 and 2.9, and the degree to which TAPs were tiring/not tiring was 3.4 and 3.8. Respondents reported that the degree to which TAPs were unnatural/natural was 3.4 and 3.0, and time-consuming/not time-consuming was 3.2 for both concurrent and retrospective groups. Additionally, the authors of the paper asked participants whether performing TAPs changed their normal working behavior, and found that participants thought it was only slightly different than usual. Finally, participants were asked about the presence of the TAP facilitator and recording equipment. While neutral responses were given when respondents were asked about the presence of the facilitator and/or equipment being unpleasant or unnatural, a more extreme response was given when they were asked whether it was disturbing. Specifically, the concurrent TAP group reported a mean score of 4.3 out of five for the last question. This finding was also significantly different between the concurrent and retrospective TAP groups. Aside from these findings, as of yet, little evidence appears to have been collected with respect to subjects' beliefs about their own TAPs.

2.6.4.1 First-Person Authority

Some researchers may question whether student perceptions of their own TAPs are trustworthy. One theory that is relevant for having students make judgments about their own verbalizations comes from the field of philosophy and is called "first-person authority". This is the claim that "I know better what I myself am thinking than you do" (p. 458, Levering, 2006). If

first person authority is assumed here, it should be understood that collecting student reflections of their TAPs, including the factors that aided or hindered their TAPs, is another piece of information about this type of research. In this study, the student reflections are not viewed as being any more reliable than other sources of information, however they do add to the overall picture that is presented regarding the information that TAPs can provide, including the verification of TAPs.

2.7 Applications of TAPs

2.7.1 Usability Testing

TAPs have been utilized for a variety of research purposes. However, the area of usability testing views this technique as a principal research method and, in this field, its use is generally accepted to have high face validity (van den Haak et al., 2003). Given the popularity of the practice, much of the research regarding TAPs is specific to this discipline. Though there are many findings that can be carried over into academic assessment, a distinction should be made between the research goals of educational assessment and usability testing. Certainly performance, such as the ability to successfully complete a task, is central to assessment. However, one of the main goals of usability research is to detect usability problems. That is, having a participant navigate a website or test out a new software can pinpoint circumstances that may arise that hinder the usability of the program. In some ways, academic assessment is also concerned with problems of usability. However, specific aspects of cognitive processes involved in answering an item are important to educational researchers, especially if complex thinking is meant to be utilized. Additionally, questions related to group differences and fairness may be important to educational researchers but not usability researchers. Furthermore, while TAPs for usability research aims to eradicate problems with use, TAPs for assessment purposes

is generally used for validity evidence. That is, TAPs are not meant to be a means to an end, but rather are used as another source of evidence toward an argument of validity, and an ongoing one at that.

2.7.2 Choices and Problem Solving

In addition to usability testing, TAPs can be used for research that aims to understand reasoning and problem solving techniques in non-assessment situations. For instance, Kuusela and Paul (2000) examined concurrent and retrospective protocols by posing participants with hypothetical choices for homeowner insurance policies. Fonteyn et al. (1993) describe practices of protocol analysis in the context of understanding reasoning during problem solving.

On one hand, results from this type of research may be generalizable to complex thinking tasks because complex thinking may be part of the problem solving and reasoning process. However, in most problem solving research, just as in usability testing, participants are often not school-age individuals, but rather adults. Given this sample population, and the understanding that the reasoning of children and adolescents can vary from that of adults, it is important to consider that methodological recommendations based on this type of research may not be fully applicable to TAPs in academic assessment research.

2.7.3 Academic Assessments

As described throughout this and the preceding chapter, the purpose of this research is to investigate how TAPs can be used for the validation of assessments of complex thinking. As described earlier, the *Standards* (AERA et al., 1999) list response processes as one method of collecting evidence to support validity, though, this method is generally underutilized when contrasted with other sources of validity evidence. Nonetheless, it can and has been used to build validity arguments for assessments, including assessments for constructs such as reading

comprehension (Anderson, Bachman, Perkins, & Cohen, 1991) and science performance (Ayala, Yin, Shultz, & Shavelson, 2002). This research hopes to expand the application of TAPs to include assessments of complex thinking.

2.8 TAPs as Data

After TAPs are completed, the verbalizations can be transcribed and used as data. The transcribed verbalizations are then coded, which is the interpretation model that is used to investigate the research questions (Ercikan & Roth, 2006). Depending on the intent of the research, different information may be extracted from the data. For instance, Ercikan, Gierl, McCreith, Puhan, and Koh (2004) used TAPs in their research on the equivalence of French and English versions of an assessment. The focus of their coding was specific to their research. Accordingly, verbalization transcripts were coded for evidence of students correctly answering and understanding the items, if students found specific elements of the items to be difficult, and what aspects of the items helped of confused students. The TAP data was then used to support the findings of DIF analyses.

2.9 Summary

This chapter was intended to introduce important information about TAPs regarding what has been presented in the literature, though the chapter also serves to point out critical gaps in TAP research with respect to how it is used for validation of assessments, and complex thinking assessments in particular. For instance, though various methodological findings have been published, there appears to be a limited amount of literature that is specific to TAPs for the validation of academic assessments, and certainly not assessments of complex thinking. As more school curricula advance to include complex thinking and its subsequent assessment (Common Core State Standards Initiative, 2010; Sands, 2013), validation efforts that adequately support the

uses of such assessments should evolve and be expanded upon. Given previous research on TAPs and how they can be utilized, they are promising methods of validating assessments of complex thinking because of the capability to gain insight into cognitive processes of the examinees. Motivated by the gaps and limitations described in this literature review, this investigation will pursue the research questions that were first introduced in the preceding chapter.

3 Methodology

3.1 Design Overview

This research builds on a broader research project on assessment of historical thinking, which involved administering an assessment to secondary school students individually using TAPs, as well as a large-scale administration of the same assessment. The objective of this research is to understand how TAPs can be used for validating assessments of complex thinking. The first research question examines the extent to which TAPs provide validity evidence above and beyond psychometric evidence. Sub-questions to that research question examine how TAPs provide additional validity evidence depending on item type and also how TAPs provide evidence depending on the gender of the examinee. The large-scale administration of the assessment is used to provide psychometric information about the assessment and this data is used in conjunction with the TAPs.

The second research question investigates the verification of the TAPs from examinees' perspectives. This question is addressed by examining responses that students provided following their TAP sessions. The procedures for addressing both research questions are explained in more detail in this chapter.

3.2 Measures

The two measures that were used to collect student data were a history assessment, which focused on events of World War I (WWI) and the internment of Ukrainian-Canadians in Canada during this time, and a student background questionnaire. The assessment focused on WWI because it is one of the key topics in history for Grade 11 students in British Columbia, Canada. The history assessment was a viable option for collecting TAPs because it is an academic assessment intended for high school students with a complex thinking component, and this

research is specifically interested in examining TAPs as a validation method for complex thinking assessments. The assessment is also appropriate for this research because it contains MC and CR items. The assessment length allowed a sufficient amount of data to be collected without being too time consuming, with administration of the TAPs taking between 48 to 118 minutes. The assessment and student questionnaire are provided in appendices B and C, respectively.

3.2.1 Assessment

The assessment comprised two sections. The first part was made up of 15 MC questions, which were meant to measure factual knowledge about WWI. These 15 questions consisted of content with which students would have been familiar due to recently completing a unit on WWI. For instance, the questions in this section of the test covered events leading up to the war such as the assassination of Archduke Ferdinand, and direct effects of Canada's involvement in the war.

The second part of the assessment was meant to measure students' historical thinking. The TAPs that are used to answer the research questions are from this part of the assessment. Students were given background information and then asked to read passages and sample documents related to the internment of Ukrainians during WWI. Many students indicated that they had not known about this event in Canadian history prior to learning about it from the assessment, so their success on this part of the assessment was related to their understanding and interpretation of the documents. This section was made up of 11 items, five of which were CR (i.e., short or long answer) and six of which were MC, as displayed in Figure 3.1. Each of the MC items had two score levels, four of the CR items had three score levels, and one CR item had

four score levels, as shown in Table 3.1. As students completed the assessment, they were asked to record their answers for both CR and MC items on a separate answer sheet.

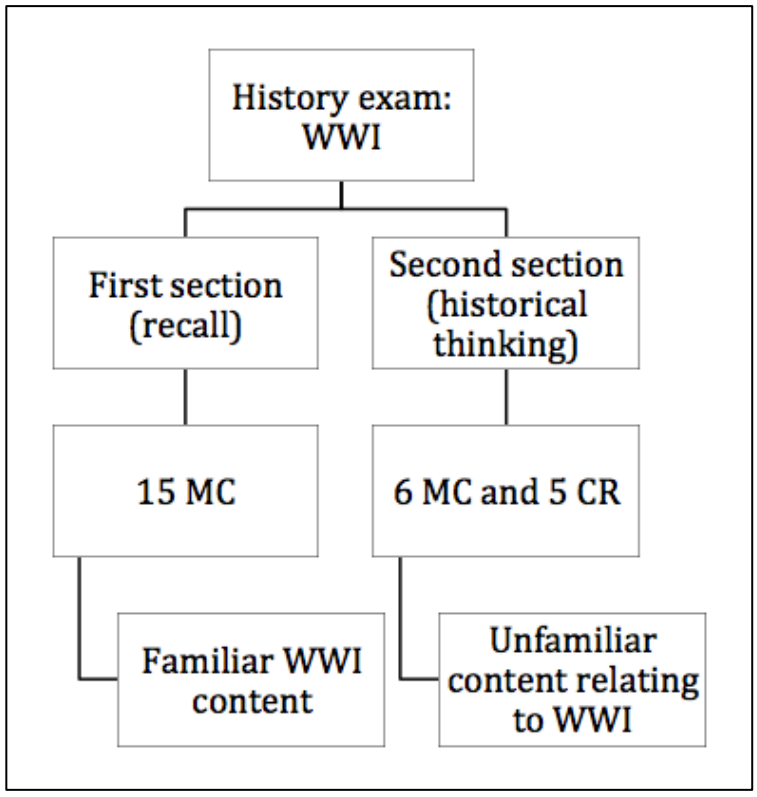

Figure 3.1 Item types and content of the assessment

Table 3.1 Item format and score levels

Item	Format	Score levels
1	MC	0, 1
2	MC	0, 1
3	MC	0, 1
4	CR	0, 1, 2
5	CR	0, 1, 2
6	MC	0, 1
7	MC	0, 1
8	CR	0, 1, 2
9	MC	0, 1
10	CR	0, 1, 2, 3
11	CR	0, 1, 2

The assessment was developed by the two professors described earlier, with the help of myself, as well as another doctoral student who has extensive knowledge of history education. The constructs that were meant to be measured by the assessment were defined and then items were written accordingly. The assessment was then pilot tested with a group of 20 students subsequent to covering WWI in their class. The pilot examinees provided feedback about the wording of questions, and subsequent changes were made.

3.2.2 Student Background Questionnaire

The other tool that was used to collect student data was a demographic questionnaire, which consisted of 34 questions in total. The first 12 questions asked students to describe themselves and their families, including age, gender, Canadian residency, cultural background, language practices, and parental education. As described earlier in the description of the sample, one question asked students to report the number of books in their home and another asked students to self-report their usual marks from their social studies classes.

The remaining questions in the questionnaire asked students about their perceptions of classroom practices. For instance, students were asked about activities that occur in their history classes, such as types of classroom discussions and using different historical resources to understand the past. Finally, students were also given an opportunity to list history-related goals from their class that the questionnaire did not cover.

3.3 Participants

3.3.1 Large-Scale Assessment Participants

Four hundred and forty one Grade 11 students from the Central Okanagan School District in Kelowna took part in the large-scale administration of the assessment. Ethics approval was

granted from the University of British Columbia (UBC), and subsequent approval was then given from the district. One lead research assistant approached Grade 11 teachers in the Kelowna school district because of his professional ties to the district. The research assistant explained the purposes of the study, and provided consent forms to be distributed to students and their parents. Upon visiting the classrooms, the research assistant collected consent forms and administered the assessment to students. Background information about the sample is described in Table 3.2. Of the 441 students, 57% (n=250) were female and 43% (n=191) were male. Eighty one percent (n=356) of the students indicated that they were 16 years old. Eighty seven percent (n=383) of the students indicated that they were born in Canada. With respect to their parents' background, 78% (n=343) said that their mothers had been born in Canada and 76% (n=334) said that their fathers had been born in Canada. Eighty eight percent (n=388) indicated that English was the most commonly used language in their home. Two other questions that were posed to students were regarding their parents' highest level of education and the number of books in their home. Regarding the question about their parents' education, 12% (n=54) of students reported that at least one parent had post-graduate schooling, 36% (n=160) of students reported a parent having a university degree, 34% (n=152) reported some college or vocational training, and 15% (n=64) reported high school or less. With respect to the question about books in the home, 47% (n=208) said that they have "Enough to fill several book cases (more than 100)", 37% (n=163) said "Enough to fill one book case (26-100)", 10% (n=46) said "Enough to fill one shelf (11-25)", and 5% (n=22) of students said "Few (0-10)". Lastly, students were asked to report the mark that they usually get on social studies tests and projects. Forty one percent (n=179) of students reported getting an A, 35% (n=153) indicated that they usually get a B, 14% (n=62) reported a C+, and 6% (n=28) said that they typically get a C. One percent (n=4) of students reported

usually getting a C-, 0.5% (n=2) of students said that they usually get an F, and 3% (n=13) of students provided multiple responses for this question.

3.3.2 TAP Participants

Thirty-five students participated in the TAP research by verbalizing their thoughts as they completed a history assessment. Students were from three Social Studies 11 classes in two schools, and all three classes had recently finished a unit relevant to the assessment. The recruitment process consisted of one principal investigator from the project communicating the nature of the research to social studies teachers at two schools in the Vancouver area. Three teachers agreed to take part in the study, and consent/assent forms were then given to the teachers who in turn distributed them to students in their classes. Students who were interested in participating in the study were asked to have the forms signed by both themselves and a guardian and returned to the teacher. Consent forms described the nature of the research to students and parents, so those who returned the signed consent forms were aware that the research involved thinking aloud as they took part in an assessment that would be administered after school. On data collection days, students who had provided their consent forms were paired with test administrators at random. Ethics approval from UBC and approval from the Vancouver School Board (VSB) had been granted prior to collecting TAPs. Ethics approval from UBC was granted by providing information about the research design, as well all assessment materials and consent forms. Approval from the VSB was subsequently granted by providing proof of ethics approval from UBC.

The characteristics of the TAP sample are presented in Table 3.2. Twenty-five students were from two classes at a secondary school in the Vancouver area and the remaining 10 students were from one class at an enriched program offered by the VSB. Of the 35 students, 10

were male and 25 were female. Most (n=24) indicated that they were 16 years old, though 11 students reported that they were "15 years old or younger". With respect to birthplace and residence, 80% (n=28) reported that they were born in Canada, and 86% (n=30) stated that they had lived in British Columbia all their lives or moved there before elementary school. Concerning their parents' background, many students (n=27) reported that their mothers had been born outside of Canada, and a comparable amount (n=26) indicated the same about their fathers. Thirty-seven percent of students (n=13) said that the most commonly used language in their home was Cantonese or Mandarin, though a similar amount (n=12) reported that English was most common. Six students indicated "Other" as the most frequently used language and 4 students reported more than one. When asked about the amount of time people in their home speak a language other than English, 37% (n=13) said "All or most of the time" and 31% (n=11) said "About half of the time". Twenty percent (n=7) and 11% (n=4) said "Once in a while" and "Never", respectively. With respect to parental education, 7 students reported that at least one parent had post-graduate schooling, 11 students reported a parent having a university degree, 10 reported some college or vocational training, 6 reported high school or less, and 1 student provided multiple answers. Answering the question of how many books their family owns, 40% (n=14) said that they have "Enough to fill one book case (26-100)" and 34% (n=12) said "Enough to fill several book cases (more than 100)". Seven students said "Enough to fill one shelf (11-25)" and just 2 students said "Few (0-10)". Finally, students were also asked to report the mark that they usually get on social studies tests and projects. Almost half of the students (n=17) reported getting an A, 12 said that they usually get a B, 2 said C+, and another 2 said C. No students reported getting less than a C, though 2 provided multiple marks.

Table 3.2 Demographic information for TAP and large-scale assessment samples

	TAP sample	Large-scale assessment sample
Gender		
Male	10 (29%)	191 (43%)
Female	25 (71%)	250 (57%)
Age		
15 years old or younger	11 (31%)	47 (11%)
16 years old	24 (69%)	356 (81%)
17 years old	-	35 (8%)
18 years old	-	3 (1%)
Place of birth (Canada)		
Student	28 (80%)	383 (87%)
Mother	8 (23%)	343 (78%)
Father	9 (26%)	334 (76%)
Most commonly used language at home		
English	13 (37%)	388 (88%)
Mandarin or Cantonese	12 (34%)	2 (1%)
Highest level of schooling that either parent attended		
High school or less	6 (17%)	64 (15%)
Some college	10 (29%)	152 (35%)
University degree	11 (31%)	160 (36%)
Post-graduate	7 (20%)	54 (12%)
Number of books in the home		
Few	2 (6%)	22 (5%)
Enough to fill one shelf	7 (20%)	46 (10%)
Enough to fill one bookcase	14 (40%)	163 (37%)
Enough to fill several bookcases	12 (34%)	208 (47%)
Typical social studies grade		
A	17 (49%)	179 (41%)
B	12 (34%)	153 (35%)
C+	2 (6%)	62 (14%)
C	2 (6%)	28 (6%)
C-	-	4 (1%)
I	-	2 (1%)

Note. Percentages are rounded to the nearest whole number.

3.3.3 Comparison of the Two Samples

The comparison of psychometric properties of an assessment with information about cognitive processes is central to the first research question in this study. Therefore, it is important to examine and acknowledge similarities and differences between the large-scale administration sample, which is used to examine the psychometric properties of the assessment, and the TAP sample, which is used to examine the cognitive processes of students participating in the assessment. The two samples come from separate school districts and have some demographic differences between them. Both samples have larger groups of female than male subjects, but the proportion is larger for the TAP sample (71% compared 57%). The large-scale administration sample tends to be older students with a larger proportion of 16-year old students (81% compared to 69%) and smaller proportion of 15-years old students (11% compared to 31%). The large-scale administration sample also includes 17- and 18-year old students (8% and 1%, respectively), whereas the TAP sample does not. Both samples have a very similar proportion of students who were born in Canada (80% for the TAP sample and 87% for large-scale administration sample). However, most of the TAP sample has parents who were born outside of Canada, compared to the large-scale administration sample who mostly reported having parents born within Canada. Language spoken at home was also different for the two samples: a great majority (88%) of the large-scale administration sample reported English to be most often spoken at home, but only 34% of the TAP sample reported the same answer. Notably, 37% of the TAP sample reported that Cantonese or Mandarin was most often spoken at home, compared to just 1% of the large-scale assessment sample. Educational background of parents was similar for the two samples, with 51% of the TAP sample and 48% of the large-scale administration sample reported having at least one parent who has a university degree or higher.

The number of books at home was included in the background questionnaire as a pseudo socio-economic status indicator. There were small differences (84% for the large-scale administration sample versus 74% for the TAP sample) in the proportions of students who reported having at least one bookcase full of books at their homes. With respect to students' reported grades, the two samples were fairly similar. For example, 49% of the TAP sample reporting that they usually receive an A compared to 41% of the large-scale assessment sample. Thirty-four percent of the TAP sample reported that they usually receive a B, compared to 35% of the large-scale assessment sample, and 12% of the TAP sample reported that they usually receive a C or C+ compared to 20% of the large-scale assessment sample. As noted earlier, of the 35 students from the TAP sample, 10 of those students were part of an enriched program offered by the VSB.

3.4 TAP Administrators

Test administrators were comprised of eight graduate research assistants and two professors who were the principal investigators of the project. Each data collector administered between two and seven TAP sessions. Four of the graduate research assistants were measurement students and four were students in curriculum and pedagogy. Three of the curriculum and pedagogy students were primarily interested in history education. The remaining curriculum and pedagogy student was primarily interested in investigating immigrant and international educational experiences through mixed methods, including TAPs. Of the two professors, one was an expert in measurement and one was an expert in historical thinking. Of the ten test administrators, four were male and six were female, and all were fluent English speakers. The ages of administrators ranged from their late 20's to early 60's. Test administrators were recruited if they had worked directly with one of the two principal investigators in the past, and had indicated their interest in the type of research that was involved in the project.

3.5 Procedure

3.5.1 Large-Scale Assessment Administration

As described earlier, one lead research assistant administered the large-scale assessment. The research assistant took approximately 10 minutes to describe the test and then students were given the rest of the 80 minute class period to take the assessment and provide their answers for the student background questionnaire. The assessments and the questionnaires were group administered in classrooms.

3.5.2 Training of TAP Administrators

Prior to conducting the TAPs, test administrators met to discuss and practice administering the sessions. The lead researchers discussed the purpose of the research to the administrators, and explained the specific role that TAPs had in the research goals of the project. Test administrators were told that they would be administering the history assessment to high school students, and were also introduced to the materials that they would be using to collect the data. As a lead research assistant on the project, I explained best practices with respect to TAP data collection.

Each administrator had a chance to take the assessment and simultaneously verbalize his or her thoughts while another practiced administration, including taking notes on the administrator notes sheet, which is described below. Test administrators were reminded to make an effort to minimize their interruptions and use prompts such as “keep thinking aloud”, and that the entire session was recorded using digital audio recorders that were provided. After practicing, administrators discussed the session and asked any questions that came up during administration. The training session also included a discussion of the order of which events were to occur, such as when to administer the demographic questionnaire and present the honorarium.

3.5.3 Introduction and Warm Up Exercise

For TAP data collection, administrators met at the school where the research was taking place and I distributed materials, including copies of the assessment, questionnaire, additional forms, digital audio recorders, and honoraria. After the last period of the school day, administrators went to the classroom of the teacher who was taking part in the research, and met those students who had returned their consent forms. Students were randomly paired with a test administrator and they were given an empty classroom in which to conduct the session. Students were instructed to sit at a desk or table and the test administrator sat either beside or across from the student depending on preference and convenience. Upon sitting down for the administration of the TAPs with student participants, the researcher introduced himself or herself and described the nature of the testing session. The data collector told the participant his or her name and that (s)he was a researcher from the University of British Columbia. The purpose of the research was described, with an emphasis that the exercise was not meant to measure student knowledge about history. A description of the assessment was given, and finally, students were informed that the results from their assessment would be kept confidential and not be shared with teachers, parents, or other students. Consent and assent forms signed by the student and their parent had been previously signed and handed in to the students' history teacher. The researcher then read the following instructions to the student:

"I would like you to start reading the questions aloud and tell me what you are thinking as you read the questions. After you have read the question, interpret the question in your own words. Think-aloud and tell me what you are doing. What is the question asking you to do? What did you have to do to answer the question? How did you come up with your answer? Tell me everything you are thinking while you are responding to the question. Let's try a practice question before we start. I'll go first. I'm going to read the passage and then answer the first question. (After administrator models the TAP): Now you read the passage and answer the second question"

Test administrators then introduced a warm up exercise in which the student watched the data collector demonstrate thinking aloud while reading a passage and answering a question. The passage was not related to WWI, but rather an excerpt from a recent newspaper article regarding politicians' views on the Canadian gun registry, and the corresponding MC item was related to the passage. After the researcher had modeled the TAP, the student read the same passage and answered another MC item related to the passage while thinking aloud. It was during this time that the researcher could gauge the student's understanding of the task, provide further directions, and then choose to move on to the actual assessment if the student was ready.

3.5.4 TAP Administration

Once the student was set to begin, s(he) was given two documents: the assessment and accompanying answer sheet to record his or her answers. All sessions were recorded with digital audio recorders. The data collectors could prompt participants to "*keep thinking aloud*" if they observed that the student fell silent while completing the assessment, but as described earlier, were otherwise instructed to not intervene with the assessment of the student. In the second section of the assessment, if the student had not sufficiently described his or her understanding of the question or how (s)he decided upon his/her answer, the administrator could ask, "*In your own words, tell me what the question asks*" and "*How did you come up with your answer to this question?*". TAP sessions took between 48 and 118 minutes, with a mean time of 76 minutes.

3.5.4.1 Notes Sheet

As the student worked through the test, the administrator used a notes sheet to keep track of the student's start and end time for each item, the student's answer, and whether or not the student stumbled on or misunderstood any words or concepts. While the student was completing the second section of the assessment, the administrator also used the notes sheet to jot down

specific verbalizations that students made with respect to rewording the question and strategies used to reach an answer. If students did not independently produce these types of verbalizations, the data collector could retrospectively prompt the students for this information, as described in the section above. The notes sheet had space for the administrator to note down the student's response to those questions.

3.5.4.2 Questions Regarding Verification

At the end of the TAP administration, data collectors asked the student about the similarities of their thoughts and the verbalizations that they had produced during the session. This question was meant to address one of the research questions outlined in the first chapter. Further discussion of this part of the research will be described shortly.

3.5.5 Honorarium

Finally, after the TAP students had completed both the assessment and the student questionnaire, they were thanked for their help with the research and presented with a $30 honorarium for their time. Students who participated in the large-scale administration of the assessment did not receive an honorarium.

3.6 Scoring

After all administrations of the assessment, research assistants scored the written response items according to a historical thinking scoring rubric that had been created by the primary investigators of the project and a research assistant. Scores ranged from zero to two or three depending on the question. Out of the five written response items, four were out of two points and one was out of three points.

The training session for scorers included reviewing the rubric and practicing scoring of sample cases. Responses to the open ended questions were scored independently by two scorers. Discrepancies between pairs of scores were identified and discussed by the scorers and a consensus on scores was reached. A researcher specializing in historical thinking reviewed all scores, changes were made if necessary, and the final scores were established.

3.7 Transcription of TAPs

Some of the researchers who administered the TAPs also transcribed their own testing sessions. The remaining recordings were distributed to four research assistants, who met to discuss transcription of the TAP sessions. During the meeting, the research assistants were directed to write out each session verbatim, as well as to record instances of pauses or filler words such as “um” or “ah”. Transcribers used the administrator notes to aid with their transcriptions. The transcriptions were then organized by combining TAPs into files for each item to be used for analysis.

3.8 Coding and Analyses

3.8.1 Coding of the TAPs

Some coding of TAPs had taken place prior to this study in order to answer specific research questions posed by the principal investigators of the project. Linguistic difficulty and historical thinking codes were used by the principal investigators and were also used for this research. The coding of linguistic difficulty and historical thinking was performed in pairs. Coding was first done independently and then coders compared results and then reached consensus on any disagreements. The inter-rater reliability Kappa for linguistic difficulty ranged

from 80%-100% for all items. For historical thinking codes, inter-rater reliability was 60%-70% for most items, but as high as 100% and as low as 30%-40% for some tasks.

Linguistic difficulty was applied when students showed difficulty in pronouncing a word or expression. For instance, the term Galician was used repeatedly in the assessment to refer to Ukrainians from a specific time period. Many students either mispronounced or struggled in their pronunciation of this word.

Evidence of historical thinking in student verbalizations was also previously coded. These codes are based on historical thinking concepts that were developed by the Historical Thinking Project (Peck & Seixas, 2008; Seixas, 2010). To limit the time of the assessment to one hour, three out of the original six concepts of historical thinking were assessed (Seixas, Ercikan, Gibson, & Lyons-Thomas, 2012): the *evidence*, *perspective*, and *ethical* dimensions. Historical thinking codes that identified evidence of historical thinking in student verbalizations were then created, reviewed, and modified by the principal investigators and an RA specializing in historical thinking. Depending on the type of historical thinking that individual items were meant to elicit, student verbalizations were coded for:

- *Source*; student comments on the source of the document;
- *Perspective*; student comments on perspective;
- *Purpose*; student comments on the authors' purposes;
- *Relationship*; student corroborates or contrasts source with other documents;
- *Context*; student comments on historical worldviews or contexts;
- *Fact*; student interprets a document as fact;
- *Traces*; student interprets sources as traces;
- *Human nature*; student uses generalizations to reason about the question;

- *Fair*; student states principles of ethics or fairness;
- *Distance*; student comments on temporal distance between the time of the document and now;
- *Narrative*; student considers collective responsibility in the argument for/against reparations;
- *Inference*; student considers the effects on present-day descendants in the argument for/against reparations;
- *Judgment*; student analyzes for ethical judgments at the levels of word choice, inclusion, narrative structure;
- *Collective*; builds an argument for or against the imposition of reparations (or other measures) for a historical injustice, based on considerations of collective responsibility; and
- *Descendants*; builds an argument for or against the imposition of reparations (or other measures) for a historical injustice, based on considerations of benefits and deficits to respective present-day descendants.

Fact and *human nature* are considered to be evidence of lack of historical thinking. *Fair* could be, but is not necessarily, evidence of historical thinking. Additional codes were also developed to capture student experiences of difficulty with items, and these were coded individually by the author. The student verbalizations were coded for:

- *Nervous speech*; students ask questions to the interviewer, apologize, or use delay tactics when answering an item. This can include verbalizations such as paralanguage (mumbling, throat clearing, or laughter), and comments such as “Is this OK”, “Is this the last question?”, or “I totally forgot how to do this” (Leighton, 2011b, p.16);

- *Guessing*; students indicated that responses were not based on their knowledge and were rather based on guessing strategies. While there may have been students who did not verbalize their intent to guess, the nature of the TAPs allowed for most guessing to be apparent. For example, students who guessed would often be clear about their intent to guess by saying things such as "I'm gonna guess again" or "I'm gonna take a guess and say…";
- *Not knowing*; student indicates that (s)he does not know a concept that was used in the assessment or the correct answer (for instance, "I don't know");
- *Difficulty*; student indicates that (s)he finds the item to be difficult (for example, "um, this is hard!");
- Confusion; student signifies that (s)he is confused by an item or the student clearly does not understand the question and provides a verbalization that demonstrates lack of understanding

Finally, student verbalizations were also coded for:

- *Length*; measured by word count. Student verbalizations were measured for length after the student had read the question aloud (if relevant) and ended once the student had provided his or her answer. In some cases, the TAP administrator would have asked the student about their answer. For this measure, the student's response to the administrator would not be counted in the length because that portion of the verbalization would not have occurred spontaneously. However, if the administrator had to prompt the student to think aloud before the student reached an answer, the student's subsequent verbalization was included in the length.

The code, *length*, was exploratory in nature because it was unknown whether or not longer length of the verbalizations could be indicative of increased difficulty with an item. This code will later be discussed with respect to the findings and how length compares to other aspects of items. Table 3.3 provides a summary of all of the codes that have been described here.

Table 3.3 Summary of all codes used for student verbalizations

Aspect of verbalization	Code
Shows difficulty with a specific term	Linguistic difficulty
Poses questions, apologizes, or uses delay tactics	Nervous speech
Explicitly states that answer is a guess	Guessing
Articulates	Now knowing
Expresses that an item is difficult to answer	Difficulty
Expresses confusion or is clearly misunderstanding a question or document	Confusion
Word count of verbalization	Length
Historical Thinking:	
Comments on the source of the document	Source
Comments on perspective	Perspective
Comments on the authors' purposes	Purpose
Corroborates or contrasts source with other documents	Relationship
Comments on historical worldviews or contexts	Context
Interprets document as fact	Fact
Interprets sources as traces	Traces
Uses generalizations to reason about the question	Human nature
States principles of ethics or fairness	Fair
Comments on temporal distance between the time of the document and now	Distance
Considers collective responsibility in the argument for/against reparations	Narrative
Considers the effects on present-day descendants in the argument for/against reparations	Inference
Analyzes for ethical judgments at the levels of word choice, inclusion, narrative structure	Judgment
Builds an argument for or against the imposition of reparations (or other measures) for a historical injustice, based on considerations of collective responsibility	Collective
Builds an argument for or against the imposition of reparations (or other measures) for a historical injustice, based on considerations of benefits and deficits to respective present-day descendants	Descendants

3.8.2 Investigation of Each Research Question

The focus of this research is to understand how TAPs provide validity evidence for assessments of complex thinking. The following sections specifically outline how each research question was investigated.

3.8.2.1 Information Provided by TAPs Beyond Psychometric Evidence

The main focus of this research is to understand how TAPs provide validity evidence above and beyond psychometric evidence for assessments of complex thinking. Psychometric evidence was provided by the data collected from the large-scale administration of the assessment. The IRT difficulty and discrimination parameters were determined for each item using the software program Pardux (Burket, 1998). These parameters were estimated using the 3-parameter logistic (3-PL) model for MC items and the 2-parameter partial credit (2-PPC) model, also known as the generalized partial credit model, for CR items. The 3-PL model was used for MC items because it takes into consideration the possibility that students may guess their answer, but also estimates difficulty and discrimination parameters. The 2-PPC model was used for CR items because it estimates item discrimination and item score level difficulty parameters for items with multiple score levels. Both of these IRT models are commonly used and considered to be among the most accurate models for estimating MC and CR items (CTB/McGrawHill, 2008; Sykes & Yen, 2000). As described in the literature review, the difficulty parameter is the value that describes the likelihood that a respondent will correctly answer the item. The discrimination parameter is the ability of the item to discriminate between students of different ability levels.

3.8.2.1.1 Consistency between Item Difficulty and TAPs

The difficulty parameter was determined for each item and then compared to the TAPs, which were coded for signs of linguistic difficulty, nervous speech, evidence of guessing,

whether the student expressed difficulty with the item, not knowing, and confusion. If students are more likely to express these in their verbalizations for items with higher difficulty parameters, this would indicate consistency between the item difficulty parameter and TAP evidence. That is, an item with a high difficulty parameter should have a higher proportion of students who express these aspects of difficulty. An item with a low difficulty parameter should have fewer students who have these characteristics in their verbalizations. Length of the verbalization was also compared to item difficulty, though as described above, this code is exploratory.

3.8.2.1.2 Consistency between Item Discrimination and TAPs

In order to investigate how information from student verbalizations compares to the discrimination parameter, verbalizations of students with high and low historical thinking (HT) scores were compared for evidence of HT. The IRT discrimination parameter is an item characteristic that indicates how well that item discriminates between high and low ability students. Differences in evidence of HT in student verbalizations for high and low HT students may corroborate the information that is provided by the discrimination parameter. This would be supported if items with high discrimination parameters showed greater differences in evidence of HT in student verbalizations compared to items with low discrimination parameters.

Students' HT score was determined by their total score on the portion of the assessment that was targeted to assess complex thinking. The HT scores ranged between five and sixteen. A frequency distribution of the HT scores showed that using cut-off points of a) nine points or lower and b) 13 points or greater would result in three approximately equal groups. The sample was divided into three groups of high, moderate, and low HT groups based on these cut-scores, with sample sizes of 11, 12, and 12, respectively. The moderate group was used to adequately

separate the high and low HT students. For the investigation of TAPs and their relationship to item discrimination, only the high and low HT groups were used.

3.8.2.1.3 Consistency between Factor Analysis and TAPs

An exploratory factor analysis (EFA) was performed in order to examine the factor structure and factor loadings of the items based on the large-scale assessment response data. The factor structure signals the presence of underlying dimensions, that is, the number of constructs that are being measured by the items in the assessment. Factor loadings indicate how well the individual items measure those constructs. The results from the EFA were compared to evidence of HT in student verbalizations, which was described earlier in the coding section of this chapter.

It should be noted that a factor analysis examines the assessment at the test level, while the coding of TAPs examines the assessment at the item level. However, because the factor analysis is based on inter-item correlations, similarities among items based on verbalizations can provide insights about underlying dimensions that a factor analysis may reveal. Another possibility is that the dimensions from a factor analysis may be influenced by item characteristics aside from the construct, such as item format like MC versus short answer. In such a case, verbalizations may be useful in examining and documenting defining characteristics of the construct that a factor analysis may fail to capture.

3.8.2.1.4 Gender Differences

A sub-question, related to the overarching research question described above, investigates how TAPs provide validity evidence about gender groups. This question aims to look specifically at the differences in TAPs from male versus female students and whether those TAPs provide additional evidence to psychometric information about different response patterns between the gender groups. DIF analyses were performed on the large-scale data using the IRT-

based Linn-Harnisch (LH) DIF detection method (Linn & Harnisch, 1981) with the program, Pardux (Burket, 1998). The LH method analyzes both dichotomous and polytomous items. The Mantel-Haenzsel (MH; Holland & Thayer, 1988; Mantel & Haenszel, 1959) and Mantel (Mantel, 1963, Zwick, Donoghue & Grima, 1993; Zwick, Thayer & Mazzeo, 1997) DIF detection methods, for dichotomous and polytomous items respectively, were also performed using the DIFAS program (Penfield, 2007).

The LH method determines if gender DIF is present by estimating the parameters for the entire group and then determines the fit of the item parameters for the focal group. The z-score for fit, combined with the difference between the observed and expected means, indicates the level of DIF that is present. A z-score $\geq |2.58|$ and a difference between the observed and expected probability of correct response $\geq |0.1|$ indicates level 3 DIF, which is the most extreme case of DIF. A z-score $\geq |2.58|$ but has a difference in observed versus expected mean $\leq |0.1|$ indicates level 2, or moderate, DIF. Level 1 DIF, which means that DIF is not present, is indicated by a z-score $\leq |2.58|$.

The MH and Mantel methods determine if DIF is present by using a chi-squared test of independence between two groups matched in ability. ETS classification of DIF indicates that an item has negligible DIF (Type A) if the index of differential performance is not significantly different from 0 and $\leq |1|$. An item is said to have moderate DIF (Type B) if the index of differential performance is significantly different from 0 and $\geq |1|$ but $\leq |1.5|$. Finally, an item is said to have large DIF (Type C) if the index of differential performance is $\geq |1.5|$ and significantly different from 1 (Hidalgo & Lopez-Pina, 2004; Zwick & Ercikan, 1989).

In this study, more than one method of DIF detection is used to verify DIF status of items given expected differences between DIF detection methods (Hambleton & Rogers, 1989;

Hidalgo & Lopez-Pina, 2004; Rogers & Swaminathan, 1993). Using the MH or Mantel technique, along with the LH method, is advantageous because it requires only a modest sample size (Sireci, Patsula, & Hambleton, 2005). Gender DIF may be present for any number of reasons. In past DIF research, MC items have been shown to favour male examinees while CR items have been shown to favour female examinees (Liu & Wilson, 2009). Other patterns of gender DIF have been found with respect to test content. For instance, Zwick and Ercikan (1989) found that DIF items that had to do with war or asked about dates of a historical event tended to favour male students. Female students were more likely favoured when DIF items were associated with civil rights, such as segregation or voting rights of women. The authors found that subgroups of examinees were more likely to be favoured for DIF items if the question was of special interest to that particular group. DIF analyses indicate which items function differentially, but they do not provide information about sources of DIF. As discussed in previous chapters, TAPs can be used as a method of explaining sources of DIF. In this study, the TAPs of male and female students were compared by linguistic difficulty, nervous speech, evidence of guessing, whether the student expressed difficulty with the item, not knowing, confusion, and length of verbalizations. These results will be compared to the DIF results, and may be used as direct evidence to show DIF and sources of DIF. For instance, if one group has higher incidences of nervous speech, guessing, expressing difficulty, and confusion for an item, one conclusion would be that the group had trouble with that particular item.

In some cases, DIF may not be evident if there is a consistent difference in difficulty between the two groups throughout the test. Therefore, item p-values for male and female examinees were compared. Regardless of whether or not DIF was present, TAPs from both gender groups were compared as described above.

3.8.2.1.5 Item Format

The second sub-question examines how student verbalizations vary on MC and CR (short/long answer) items with respect to complex historical thinking, and how that information compares to psychometric evidence about what these items are capturing. Though the TAPs from the assessment had already been examined for elements of HT (Ercikan et al., 2012; Seixas et al., 2012), the TAPs were not previously compared based on item type. Consequently, this was done to investigate whether there is a difference in complex thinking (specifically HT) based on item format (selected response versus extended response). In addition, item format was compared based on linguistic difficulty, nervous speech, evidence of guessing, not knowing, confusion, and difficulty in order to investigate if students experienced more difficulty with one item type.

3.8.2.2 Student Reported Verification

The final research question examines verification of TAPs by asking the extent to which students believed that their TAPs reflected their thinking processes. As discussed in the literature review, there has been research that has examined participant experiences of providing TAPS, and the effect that TAPs had on their performance. However, there does not appear to be any research that has examined the consistency between a person's thoughts and their TAPs, as evaluated by the person him- or herself.

Immediately after completing the assessment, the administrators posed two questions to participants. The first question asked, "When you thought aloud *while* you were answering the questions, how similar was what you said to what you were actually thinking?" Students were then asked to choose between four options: Extremely similar, Somewhat similar, Not very similar, or Not at all similar. The answers to these two questions were then examined to gain a general idea of respondents' overall beliefs about their TAPs. This was followed by comparisons

between groups of students. That is, female versus male responses will be compared, as well responses from high versus low achieving students.

Upon providing their answer to the two questions, the administrator was directed to explore why they had answered the way that they did. For instance, students who indicated that their verbalizations were "somewhat similar" to their actual thoughts were asked to consider possible explanations for why they were not exactly the same. This question was open ended so as to not limit student answers. This data was categorized and summarized as part of examining verification of TAPs.

3.8.3 Summary of Psychometric Evidence and Corresponding TAP Evidence

As discussed in the previous sections, a number of psychometric analyses will be compared to various types of TAP evidence. Table 3.4 provides a summary of each psychometric method that was used in this study, and the corresponding TAP evidence to which it was compared.

Table 3.4 Psychometric evidence and corresponding TAP evidence

Psychometric evidence	TAP evidence
IRT item difficulty	Linguistic difficulty
	Nervous speech
	Guessing
	Difficulty
	Not knowing
	Confusion
	Length
IRT item discrimination	Historical thinking
Factor analysis	Historical thinking
Gender DIF	Linguistic difficulty
	Nervous speech
	Guessing
	Difficulty
	Not knowing
	Confusion
	Length
	Historical thinking

4 Results

This chapter presents a detailed description of the results of the current study. An assessment of HT was administered to a large sample of grade 11 students in order to obtain psychometric information about the assessment. A separate, smaller student sample was administered TAPs as they completed the assessment. The data from each group were then compared to one another in order to answer the first research question, which asks how TAPs provide validity evidence above and beyond psychometric evidence for assessments of complex thinking. After completing the assessment, the TAP sample was asked if they believed that their verbalizations were accurate reflections of their thinking, and were then asked to provide reasons that caused similarities or dissimilarities. Their responses were used to address the second research question, which examined how students viewed the accuracy of their verbalizations as indicating their thought processes.

The chapter begins with a summary of descriptive statistics of the data from both the large-scale and the TAP administration of the assessment, a description of the psychometric analyses and findings, and a description of the validity evidence gathered from TAPs. The results of the analyses addressing the first research question are then presented. The validity evidence provided by TAPs is compared to psychometric information based on the large-scale administration of the assessment including item difficulty p-value statistics, IRT difficulty and discrimination parameters, factor analysis results, and DIF results. The chapter goes on to present the results for the second research question. Students reported on how their verbalizations compared to what they had actually been thinking, and then provided possible reasons for similarities or dissimilarities between verbalizations and their thought processes.

4.1 Descriptive Statistics, Psychometric Analyses, and Validity Evidence from TAPs

4.1.1 Length of Verbalizations

The first descriptive aspect of the research is the mean length of verbalizations for each of the items. Once student verbalizations were transcribed, a word count was taken for each item, up until the TAP administrator prompted the student to verbalize his or her thoughts retrospectively. That is, verbalization length was considered to be the spontaneous verbalization of the student, including verbalizations after a student was reminded to "*keep thinking aloud*", until the answer for the item was provided. As described in the previous chapter, this aspect of student verbalizations was exploratory in that previous research has not investigated if verbalization length of TAPs is indicative of ease or difficulty of a particular item. The mean length of the verbalization (in words) for each item, including those for the entire group, males, and females, low HT students, and high HT students are displayed in Table 4.1. The table also includes the mean verbalization length across all items. The entire group had an average verbalization length of 178 words per item, with a minimum verbalization length of 113 and a maximum verbalization length of 294. This table also shows that males and females had similar verbalization lengths overall, though differed for individual items. However, significant differences between males and females were not observed for any individual items. The low and high HT groups varied with respect to verbalization length, with the high HT group providing longer verbalizations than the low HT group. Statistically significant differences ($p<0.05$) were observed for items 5, 10, and 11, with Cohen's *d* effect sizes of 1.78, 1.26, and 0.85 respectively. These effect sizes are each considered to be large (Cohen, 1988).

Table 4.1 Mean verbalization length in number of words for each item

Item	All	Gender		HT	
		Males	Females	Low	High
1	118	108	122	108	110
2	151	141	155	155	136
3	204	189	210	197	208
4	184	164	191	154	223
5	124	115	128	85	178
6	165	170	164	155	133
7	113	120	111	111	107
8	172	195	163	135	198
9	152	139	157	152	136
10	278	253	289	174	416
11	294	346	273	251	380
Overall	178	176	178	152	202

4.1.2 Classical Difficulty and Discrimination Indices

The easiness index p-values for each item, which are meant to be an indication of difficulty or easiness of an item, are displayed in Table 4.2. Expressed as a value between 0 and 1, the p-value is based on the percent of examinees that received a maximum score. An item that has a low p-value is considered to be difficult, and has a low probability of being answered correctly. Likewise, an item with a large p-value has a high chance of being answered correctly, and is therefore considered to be an easy item. These results show that p-values for this assessment ranged from 0.28 (item 8) to 0.83 (item 2).

The corrected item-total correlation is a value that is commonly used as a measure of item discrimination (Furr & Bacharach, 2008). This value is the correlation between an item and the other items in the assessment, excluding the item that is being correlated. If the corrected item-total correlation is positive, it is consistent with other items in the assessment, with a greater value indicating higher consistency. Table 4.2 shows that all of the corrected item-total

correlations are positive and range from 0.21 (item 1) to 0.60 (item 10). The table also includes information about item type and score levels.

Table 4.2 Item type, score levels, and p-values, and corrected item-total correlations for each item

Item Number	Item type (MC, CR)	Max. score level	P-values	Corrected item-total correlation
1	MC	1	0.76	0.21
2	MC	1	0.83	0.36
3	MC	1	0.57	0.35
4	CR	2	0.45	0.40
5	CR	2	0.43	0.41
6	MC	1	0.74	0.35
7	MC	1	0.74	0.30
8	CR	2	0.28	0.38
9	MC	1	0.79	0.27
10	CR	3	0.48	0.60
11	CR	2	0.53	0.47

4.1.3 Exploratory Factor Analysis

A maximum likelihood exploratory factor analysis (EFA) with promax rotation using a polychoric correlation matrix was performed. In practice, a number of criteria are used to determine the total number of factors present, including eigenvalues greater than one, the slope of the scree plot, and parallel analysis (Russell, 2006). While the eigenvalues greater than one criterion would suggest a two-factor model, the scree plot and parallel analysis (Figure 4.1) suggest that a one-factor model is present. Table 4.3 displays the eigenvalues and percent of variance explained based on number of factors. There appears to be a steep increase between eigenvalues for a one-factor model compared to a two-factor model, whereas there is not a steep

increase between a two-factor model compared to a 3-factor model. Given this evidence, a one-factor model is accepted[1].

The factor loadings show that CR items load the highest on the factor (see Table 4.4). The MC items make up the lower half of the loadings, with item 1 loading particularly low at 0.228.

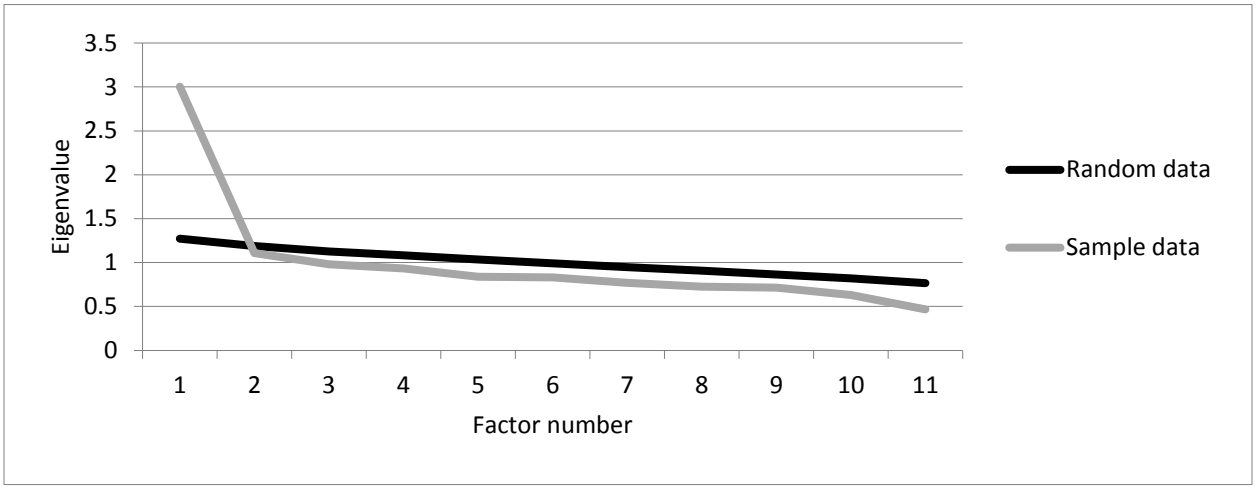

Figure 4.1 Parallel analysis showing a one-factor model

Table 4.3 Eigenvalues and total variance explained

Number of factors	Eigenvalue	Percent of variance explained
1	3.005	27.317
2	1.108	10.071
3	0.982	8.924
4	0.932	8.473
5	0.840	7.634
6	0.831	7.553
7	0.769	6.987
8	0.724	6.579
9	0.714	6.489
10	0.630	5.732

[1] The factor loadings for a two-factor model are presented in Appendix C

Table 4.4 Factor loadings from the one-factor model

Item	Factor loading	Item type
10	0.730	CR
11	0.579	CR
5	0.481	CR
4	0.469	CR
8	0.461	CR
6	0.399	MC
3	0.398	MC
2	0.389	MC
7	0.344	MC
9	0.308	MC
1	0.228	MC

4.1.4 IRT Parameters

A common assumption for most IRT models is that unidimensonality is met (Hambleton, Swaminathan, & Rogers, 1991), and this assumption has been established from the EFA that was presented in the previous section. The discrimination and difficulty parameters were determined for each item based on IRT calibration. As described in the previous chapter, parameters for MC items were estimated using the 3-PL model and parameters for CR items were estimated using the 2-PPC model. Under the 3-PL model, discrimination parameters are characterized by a-parameters, difficulty parameters are characterized by b-parameters, and a third parameter, the guessing parameter, is characterized by c-parameters. Under the 2-PPC model, discrimination parameters are characterized by alpha, and difficulty parameters are characterized by gamma (Burkett, 1998). Because the CR items have score levels that are greater than two (i.e., consist of score levels other than 0 and 1), the item score level difficulty parameter is provided for the highest score level. The 2-PPC model does not account for guessing and so a third parameter is not provided. The IRT model used for parameter estimation, as well as parameters for all of the items, are displayed in Table 4.5.

An item is more likely to discriminate between students of high and low ability if the discrimination parameter is high. That is, a student with high ability is much more likely than a student with low ability to correctly answer an item with a high discrimination parameter. Most items had moderate discrimination parameters that ranged between .5 and 1. Item 1, a MC item, had the lowest discrimination of .35, and items 10 and 11, both CR items, had the highest discrimination parameters of 1.40 and 1.13, respectively. A difficulty parameter value increases with the difficulty of an item. Therefore, an item with a low difficulty parameter is said to be an easy item, whereas an item with a high difficulty parameter is said to be a difficult item. A number of items had moderate difficulty levels with IRT difficulty parameters between -1 and 1. Item 1 was the easiest item, with a low difficulty parameter of -1.52, and item 10 was the most difficulty item, with a difficulty parameter of 2.55. Again, only the MC items had guessing parameters, because these items were estimated using the 3-PL model. The guessing parameters were all 0.20 for each item.

Table 4.5 IRT model information for each item

Item	IRT model	Discrimination parameter	Difficulty parameter	Guessing parameter
1	3-PL	0.35	-1.52	0.20
2	3-PL	0.89	-1.22	0.20
3	3-PL	0.75	0.12	0.20
4	2-PPC	0.69	0.24	-
5	2-PPC	0.81	1.45	-
6	3-PL	0.73	-0.76	0.20
7	3-PL	0.56	-0.94	0.20
8	2-PPC	0.86	0.51	-
9	3-PL	0.51	-1.35	0.20
10	2-PPC	1.4	2.55	-
11	2-PPC	1.13	1.06	-

4.1.5 Gender DIF Analysis

DIF analysis is used to identify if individual items have differential response patterns for groups, despite matched ability levels of students in those two groups. For instance, an item that is identified as favouring male students means that male students are more likely to perform well on that item relative to female students with comparable ability levels. While this type of analysis can identify which items are functioning differentially, it does not provide explanations or reasons for group differences. Examining verbalizations of students from TAPs can provide information about how students from different groups are interpreting and responding to the tasks, and thus provide insights into why students with similar ability levels may be performing differentially.

As discussed previously, the LH, MH, and Mantel DIF detection methods were used to identify gender DIF. The LH method, which can detect uniform and non-uniform DIF, was performed twice so that both males and females could serve as the focal group when compared to the combined sample. The first analysis used female students as the focal group and the second analysis used male students as the focal group. While no items were found to exhibit DIF when females were the focal group, item 4 was found to function differentially favouring females when males were the focal group because the z-score for this item was -3.119. Recall that a z-score greater than $\left|2.58\right|$ suggests that DIF is present for an item. This CR item asked students to read a passage and answer why an American government representative would describe Ukrainian prisoners differently than a priest from an earlier passage. This item was found to exhibit level 3 DIF, as opposed to a lower level of DIF, because the observed minus predicted mean for this item was -0.157. As discussed in an earlier chapter, a value greater than $\left|0.1\right|$ for

the observed-expected mean suggests level 3 DIF. The z-scores, observed, expected, and observed-expected means for all items are displayed in Table 4.6.

Table 4.6 Pardux DIF results using males as the focal group

Item	Z-score	Observed mean	Predicted mean	Observed-Predicted mean
1	0.387	0.76	0.74	0.016
2	0.779	0.82	0.8	0.018
3	0.246	0.57	0.57	0
4	-3.119*	1.72	1.88	-0.157**
5	-1.39	1.79	1.84	-0.058
6	0.925	0.74	0.72	0.024
7	1.505	0.77	0.73	0.044
8	0.227	1.55	1.56	-0.006
9	-0.528	0.75	0.77	-0.014
10	-2.107	2.31	2.41	-0.095
11	-1.687	1.97	2.04	-0.067

Note. *z-score> $\left|2.58\right|$, **observed-predicted > $\left|0.1\right|$

Following these results, analyses using the MH and Mantel DIF detection methods were performed for dichotomous and polytomous items, respectively. The MH and Mantel methods do not use the combined sample as the reference group, and so the same DIF identification is expected using either males or females as the focal group. The analysis on dichotomous items was performed first, followed by the analysis on polytomous items. Females served as the focal group both times. Table 4.7 shows that DIF was not found to be present among the dichotomous items. The MH chi-square critical value is 3.84 for a Type I error rate, and Table 4.7 shows that all of the items had values below 3.84. The ETS Categorization Scheme, which classifies DIF levels into A (small), B (medium), or C (large), shows that all items are level A.

Table 4.7 MH DIF results and ETS classification for dichotomous items

Item	MH chi-square value	ETS classification
1	0.0120	A
2	0.0157	A
3	2.1248	A
6	0.0046	A
7	1.4979	A
9	2.7530	A

The Mantel DIF analysis on polytomous items identified both items 4 and 8 as exhibiting DIF. Table 4.8 shows that item 4 was found to exhibit DIF in favour of female students. The value for the Mantel chi-square statistic is 5.063, which is greater than the critical value of 3.84 for a Type I error rate of 0.05. The DIFAS program also provides the Liu-Agresti Cumulative Common Log-Odds Ratio, which is an indicator of direction of DIF. The value is -0.456, which indicates that DIF is in favour of the focal group because it is negative. The item was classified as having B-level (medium) DIF.

Table 4.8 also shows that item 8 was found to exhibit DIF in favour of male students. This item asked students to read a passage and then answer if Canada's Minister of Justice from that time period believed that there were good reasons for the internment of Austrians. The Mantel chi-square statistic value was 9.181, which is larger than both the critical value of 3.84 for a Type I error rate of 0.05, as well as 6.63 for a Type I error rate of 0.01. This value for Liu-Agresti Cumulative Common Log-Odds Ratio is 0.724, which indicates that DIF is in favour of the reference group because it is positive. Item 8 was also classified as B-level (medium) DIF.

Table 4.8 Mantel DIF results and ETS classification for polytomous items

Item	Mantel chi-square value	Liu-Agresti Cumulative Common Log-Odds Ratio	ETS classification
4	5.063*	-0.456	B
5	0.013	-0.025	A
8	9.181*	0.724	B
10	0.086	-0.061	A
11	0.098	-0.066	A

The item characteristic curves (ICCs) for items 4 and 8 are displayed in figures 4.2 and 4.3. On the y-axis is the probability of getting a score of zero, one, or two for male (solid line) and female (dashed line) students depending on their ability level. The ability axis is the x-axis, and is portrayed as being between negative and positive four. The ICC for item 4 in Figure 4.2 shows that females have a higher probability of receiving a score of two regardless of their ability level because the dashed line representing females is higher than the solid line representing males for this particular score. The probability of receiving a score of one is higher for females up to an ability level of approximately zero, but for higher ability levels, males have a higher probability of receiving a score of one because at zero, the lines cross and the solid line becomes higher than the dashed line. Figure 4.2 also shows that males have a higher probability of receiving a score of zero for all ability levels because the solid line is consistently higher than the dashed line.

The ICC for item 8 in Figure 4.3 shows that males have a higher probability of receiving a score of one for all ability levels because the solid line is continuously above the dashed line, but their probability of receiving a score of two is only greater than females if their ability level is lower than approximately one. Females have a higher probability of receiving a score of zero for almost all, except for the highest, ability levels. The crossing of the two lines for each score level represents these last two exceptions.

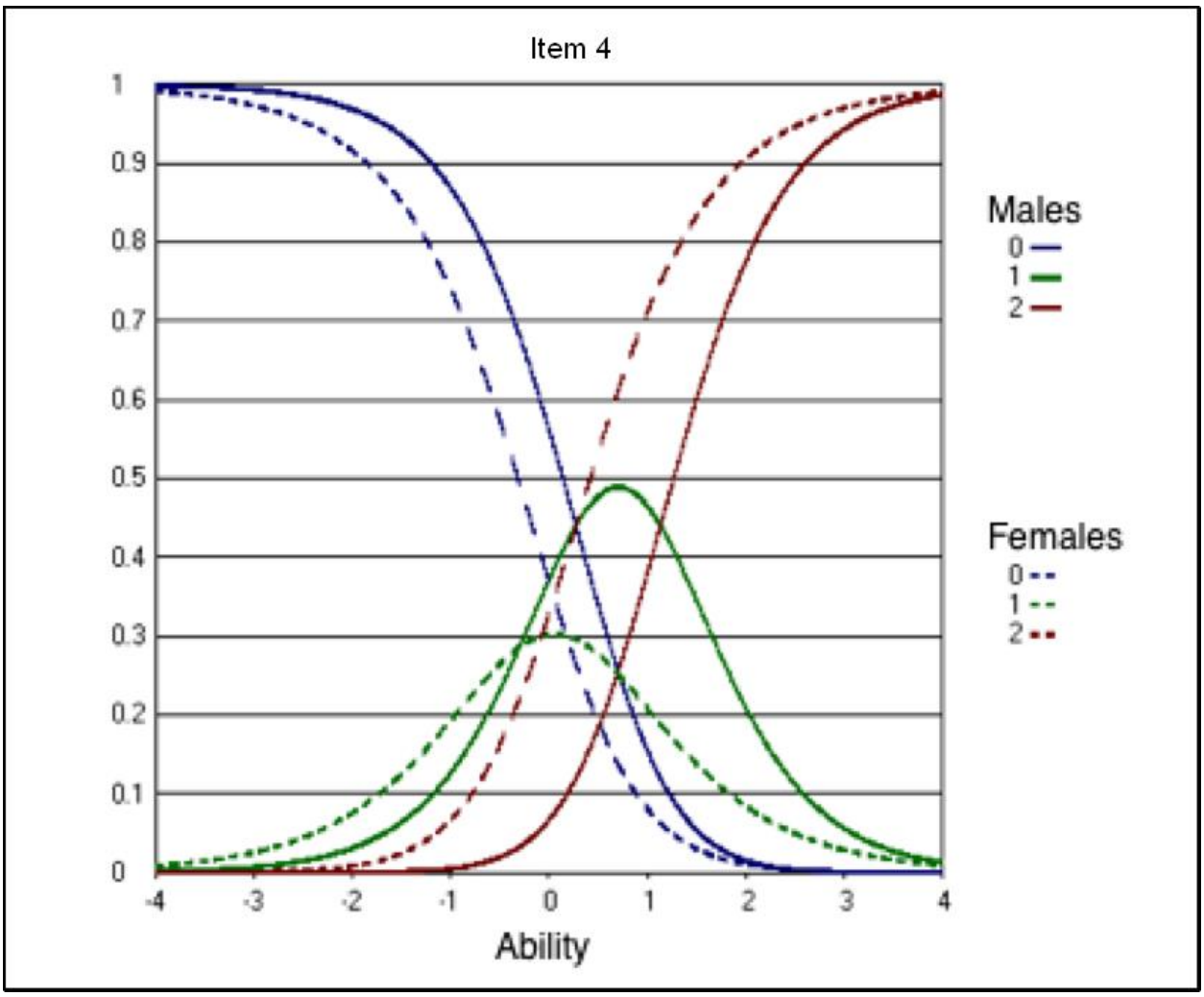

Figure 4.2 ICCs for item 4 for males and females

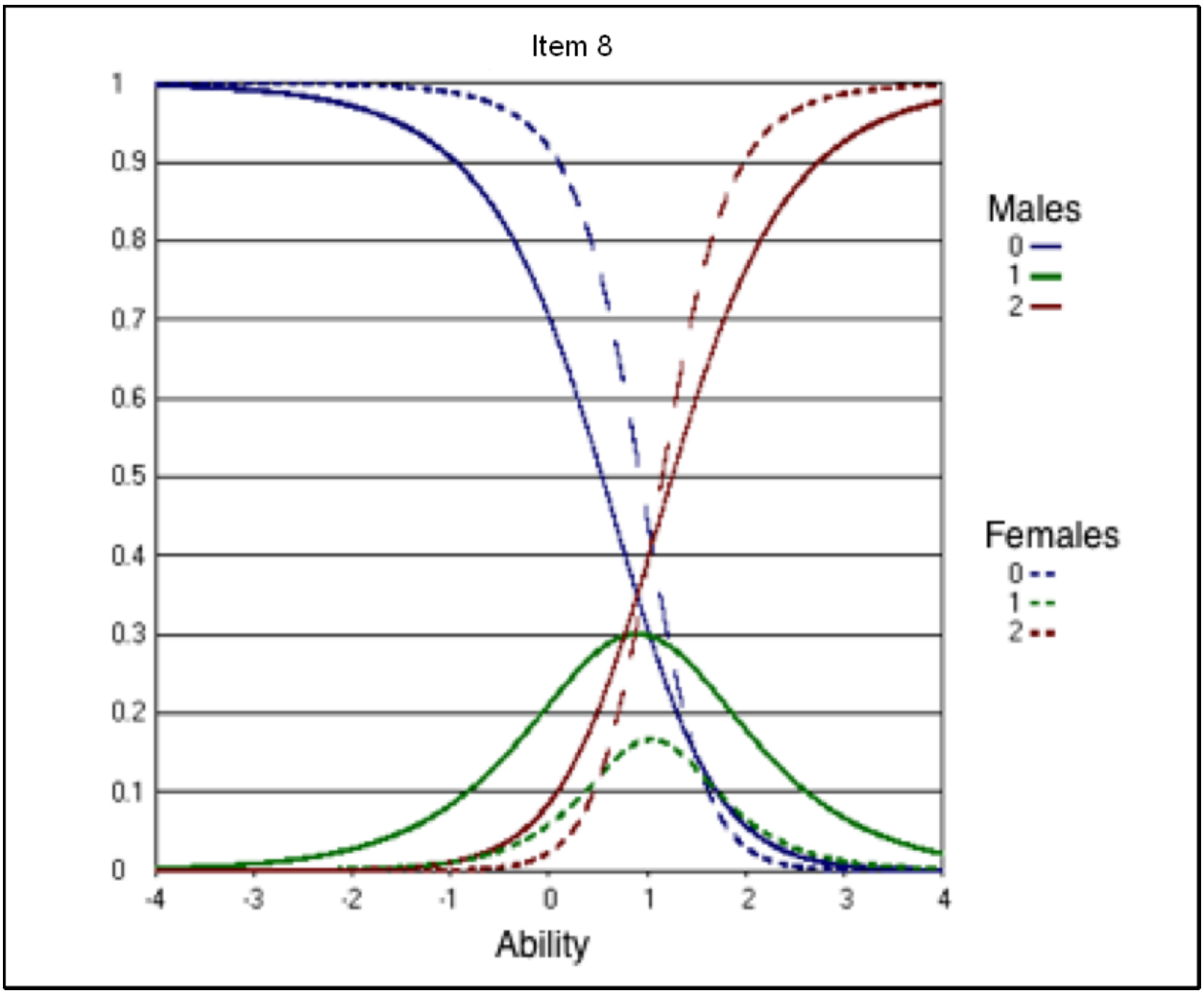

Figure 4.3 ICCs for item 8 for males and females

4.2 Validity Evidence from TAPs

After TAPs were transcribed, they were coded for evidence of HT, expressions of linguistic difficulty, nervous speech, guessing, not knowing, expressions of difficulty, confusion, and length. The codes for HT include source, perspective, purpose, relationship, context, traces, fair, distance, narrative, inference, judgment, fact, and human nature. A description of each of these codes is provided in Table 3.3 in the previous chapter.

4.2.1 Evidence of HT

TAPs provided validity evidence about how items in the assessment elicited HT. Table 4.9 shows the percent of students who provided individual aspects of HT for each item. As this table shows, not all aspects of HT were coded for each item. For instance, in coding item 1, aspects of HT that were coded were: Source, Perspective, Context, Fact, Traces, Inference, and Judgment. The table shows that, when answering item 1, 12% of students commented on the source of the document. The final column is average percentage of students including relevant historical thinking aspects in their verbalizations. This column is useful because it provides an overall indication of the extent to which students included historical thinking in their verbalizations. Item 11 stands out because it has the highest overall percent of HT (46%). The individual aspects of HT that were expected to occur for this item were: Fair, Distance, Collective, and Descendants. A high percent of students (49%, 54%, 37%, and 46%, respectively) provided evidence that each of these aspects of HT were taking place when they were answering this item.

Table 4.9 Percent of students providing evidence of HT

Item	Source	Perspective	Purpose	Relationship	Context	Fact	Traces	Inference	Judgment	Human nature
1	12%	65%	-	-	0%	21%	71%	9%	0%	-
2	21%	44%	3%	-	21%	44%	32%	9%	-	-
3	77%	17%	3%	-	60%	-	-	6%	0%	9%
4	89%	91%	17%	100%	29%	9%	17%	3%	17%	-
5	6%	43%	9%	20%	17%	-	-	57%	0%	-
6	20%	3%	-	0%	3%	29%	31%	6%	-	-
7	26%	9%	14%	-	9%	-	-	9%	0%	-
8	11%	40%	11%	9%	3%	-	3%	23%	14%	-
9	66%	43%	3%	0%	3%	-	-	9%	-	6%
10	43%	71%	3%	74%	29%	31%	57%	23%	3%	0%
11	-	-	-	-	-	-	-	-	-	-

Table 4.9 (continued) Percent of students who provided evidence of HT

Item	Fair	Distance	Narrative	Collective	Descendants	Mean HT
1	-	-	-	-	-	26%
2	-	-	-	-	-	22%
3	-	-	-	-	-	27%
4	-	-	-	-	-	45%
5	-	-	-	-	-	22%
6	-	-	0%	-	-	9%
7	-	-	0%	-	-	9%
8	3%	-	31%	-	-	15%
9	-	-	-	-	-	20%
10	37%	6%	-	-	-	35%
11	49%	54%	-	37%	46%	46%

4.2.2 Evidence of Gender Differences

The verbalizations of male and female students were compared based on expressions of linguistic difficulty, nervous speech, guessing, not knowing, difficulty, confusion, verbalization length, and evidence of HT (a mean HT score is provided based on the aspects of HT). Table 4.10 shows the percent of males and females whose verbalizations contained these various characteristics. For instance, for item 1, 20% of males showed linguistic difficulty compared to 12% of females, while no males showed nervous speech compared to 20% of females. Significant differences between genders were not found for any of these aspects of their verbalizations. However, as discussed earlier, differences in length of verbalizations were noted for certain items. These differences in length will be discussed in comparison to psychometric information about gender later in this chapter.

Table 4.10 Percent of males and females who provided various aspects of verbalizations

Item	Linguistic difficulty		Nervous speech		Guessing		Not knowing	
	Male	Female	Male	Female	Male	Female	Male	Female
1	20%	12%	0%	20%	0%	8%	0%	12%
2	10%	4%	0%	4%	0%	4%	0%	8%
3	0%	12%	0%	4%	0%	4%	0%	0%
4	10%	12%	20%	64%	0%	8%	0%	8%
5	0%	4%	40%	32%	0%	0%	0%	8%
6	10%	12%	0%	4%	10%	4%	20%	4%
7	0%	0%	0%	4%	0%	0%	0%	0%
8	20%	16%	20%	32%	0%	8%	10%	24%
9	30%	20%	0%	8%	0%	0%	0%	16%
10	10%	32%	30%	44%	0%	0%	0%	24%
11	0%	8%	30%	28%	0%	4%	0%	8%

Table 4.10 (continued) Percent of males and females who provided various aspects of verbalizations

Item	Difficulty		Confusion		Mean HT	
	Male	Female	Male	Female	Male	Female
1	0%	0%	0%	4%	33%	22%
2	0%	0%	0%	0%	22%	21%
3	0%	0%	0%	24%	28%	27%
4	0%	0%	0%	16%	46%	45%
5	0%	0%	0%	8%	23%	21%
6	0%	0%	10%	4%	10%	9%
7	10%	0%	0%	0%	17%	6%
8	0%	12%	0%	12%	18%	14%
9	0%	0%	0%	0%	18%	21%
10	0%	16%	0%	20%	27%	38%
11	0%	4%	20%	0%	43%	48%

4.2.3 Evidence about Item Type

Expressions of linguistic difficulty, nervous speech, guessing, not knowing, difficulty with items, and confusion in verbalizations were compared for the two item types. This was done in order to investigate if TAPs can identify a relationship between difficulty or unease and item format. The following tables show the mean number of times a code was identified in student verbalizations for each of the item types. For instance, linguistic difficulty was coded 22 times for MC items and 22 times for CR items. As there are six MC items and five CR items, linguistic difficulty was expressed an average of 3.67 times per MC item and 4.4 times per CR item. Table 4.11 shows average expressions for linguistic difficulty, nervous speech, guessing, not knowing, expressions of difficulty, and confusion by item format. In order to determine if the differences between CR and MC items were significant, an independent samples t-test was also performed. It was found that there was a statistically significant difference between MC and CR items ($p = 0.001$) on nervous speech. Though the sample size for this analysis was very small, with five CR items and six MC items, the Cohen's *d* effect size was 4.19, which is considered to be large (Cohen, 1988).

Table 4.11 Mean frequency of expressions of difficulty in student verbalizations for MC and CR items

Code	MC	CR
Linguistic difficulty	3.67	4.4
Nervous speech	1.83	12.8
Guessing	1	1
Not knowing	2	3.8
Expressions of difficulty	0.167	1.6
Confusion	1.5	3.2

Following student difficulty, aspects of HT were compared based on item format. Table 4.12 shows the item format averages for the various aspects of HT that were used in this research. For instance, all six MC items were coded for whether the student commented on the source of the document. This code was given 77 times for MC items, and because there are six MC items in the assessment, the average is 12.83. When MC and CR items were compared for differences in aspects of HT, statistically significant differences were not found to be present.

Table 4.12 Average frequency of evidence of HT in student verbalizations for MC and CR items

HT code	MC	CR
Source	12.83	13
Perspective	10.3	21.5
Purpose	2	3.5
Relationship	0	17.75
Context	6.6	6.75
Fact*	10.67	7
Traces	15.3	9
Inference	2.67	9.25
Judgment	0	3
Human nature*	2.5	0
Fair	N/A	10.3
Distance	N/A	10.5
Narrative	0	11
Collective	N/A	13
Benefits	N/A	16

*Note. *indicates lack of HT*

4.2.4 Evidence of Student Comprehension

TAPs showed that a number of students had trouble understanding the meaning of certain terms that were presented in the assessment. When coding verbalizations for linguistic difficulty, it was found that students had trouble with pronouncing and understanding a number of terms. Table 4.13 summarizes the number of instances of linguistic difficulty per item as well as the

terms with which students experienced difficulty. If there was more than one term identified for an item, the number of instances appears beside that term.

Table 4.13 Evidence of linguistic difficulty

Item	Number of respondents indicating linguistic difficulty	Problematic term
1	4	Prejudiced
2	2	Galician
3	3	Prejudiced
4	4	Internment (1) Ukrainian (3)
5	1	Ukrainian
6	4	Bondage (1) Ukrainian (3)
7	0	-
8	6	Doherty (3) Inspiration of sentiment of compassion (2) Hostility (1)
9	8	Dysfunctional (2) Investigating (1) Justified (1) Ukrainian (4)
10	9	Justified (7) Ukrainian (1) (Unclear; 1)
11	2	Compensation (1) Internment (1)

The TAPs identified various words with which multiple students experienced difficulty, including words that were fundamental to the assessment. For instance, when reading the word prejudiced, one student commented, “pre…..juiced….. pre….ok wait, I don’t really know what that means” while another student said, “pre-, prejud-iced, I don’t know this word.” With respect to the term Galicians, which referred to the group that was central to the assessment, one student remarked, “Galicans, Galeecans, um, I’m still not sure what that is.” The lack of understanding

of these terms, which would have undoubtedly affected a student's performance on the assessment, would not have been identified using traditional psychometric validity evidence.

4.3 The Relationship between TAPs and Psychometric Information

4.3.1 Item Difficulty

The next exploration of the data involved comparing TAP evidence to psychometric evidence of item difficulty. Item difficulty was determined using two indicators: p-values and IRT difficulty parameters. The TAP codes that were compared to p-values and difficulty parameters were linguistic difficulty, nervous speech, guessing, not knowing, difficulty, confusion, and length. These verbalizations are expected to indicate if students experience difficulty in responding to the test items.

The relationship between p-values and difficulty parameters was first examined since both are used as indicators of item difficulty and the correlation ($r = -0.773$, $p=0.005$) was high as expected. It should be noted that the correlation is expected to be negative because a high p-value indicates an easier item, whereas a high IRT difficulty parameter indicates a more difficult item.

When IRT difficulty parameters were compared to the TAP codes for each item, nervous speech ($r=0.724$, $p<0.05$), difficulty ($r=0.640$, $p<0.05$), confusion ($r=0.654$, $p<0.05$), and length ($r=0.673$, $p<0.05$) were all significantly correlated, as displayed in Table 4.14. The results indicate that students tended to express these verbalizations more when the test items were more difficult. Figures 4.4 through 4.7 show the scatterplots for the significant correlations. Using Mahalanobis distance at $p<.001$, none of the data points can be considered outliers. TAP codes are expressed as a mean value because one student's verbalizations were not recorded for the first two items, and so the sample size was 34 rather than 35 for items 1 and 2.

Table 4.14 Correlations between item difficulty and student verbalizations

	IRT difficulty parameter	p-value
IRT difficulty parameter	1	-.773**
p-value	-.773**	1
Linguistic difficulty	0.167	-0.127
Nervous speech	.724*	-.769**
Guessing	-0.285	-0.125
Not knowing	0.318	-0.408
Difficulty	.640*	-0.536
Confusion	.654*	-.637*
Length	.673*	-0.381

Note. *p<0.05, **p<0.01

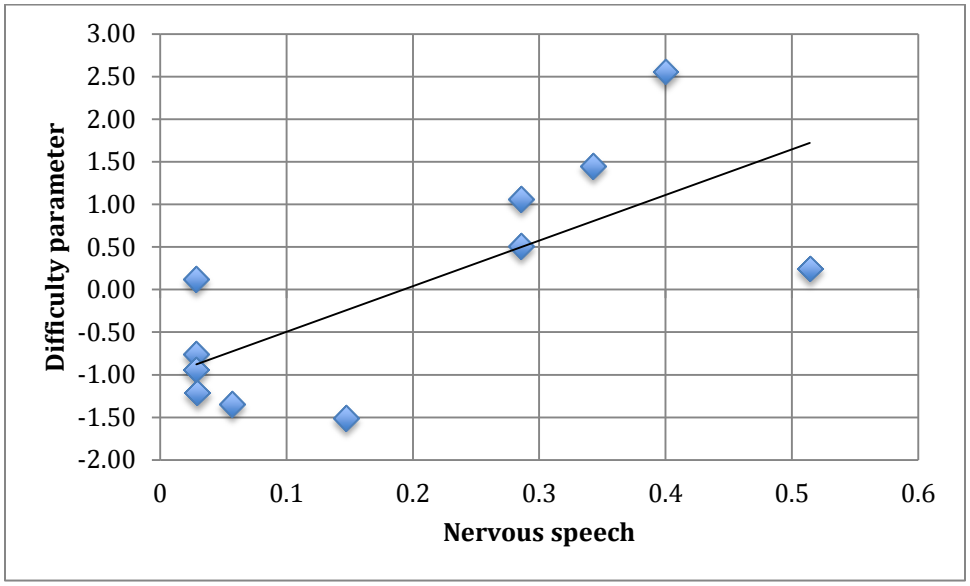

Figure 4.4 Scatterplot of difficulty parameters and verbalizations of nervous speech

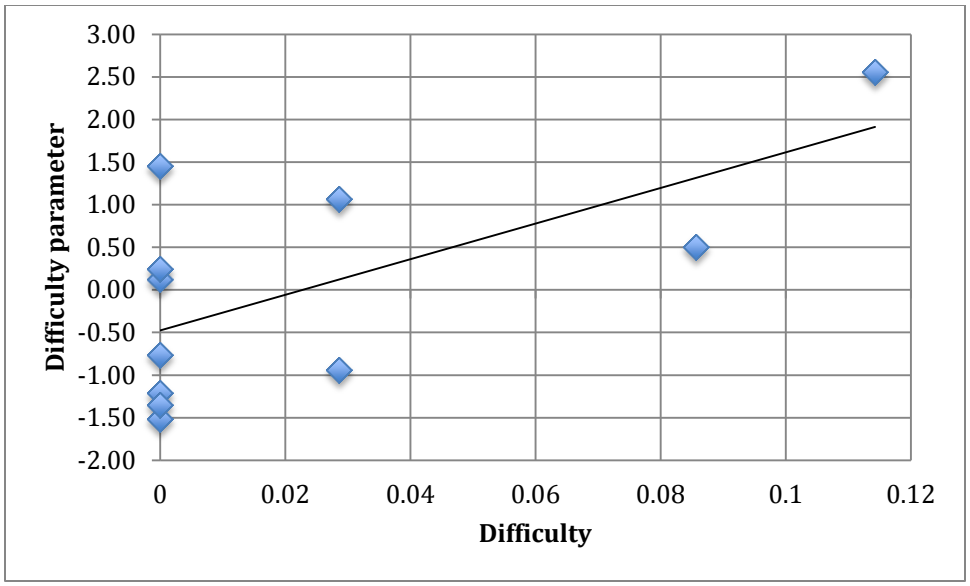

Figure 4.5 Scatterplot of difficulty parameters and verbalizations of difficulty

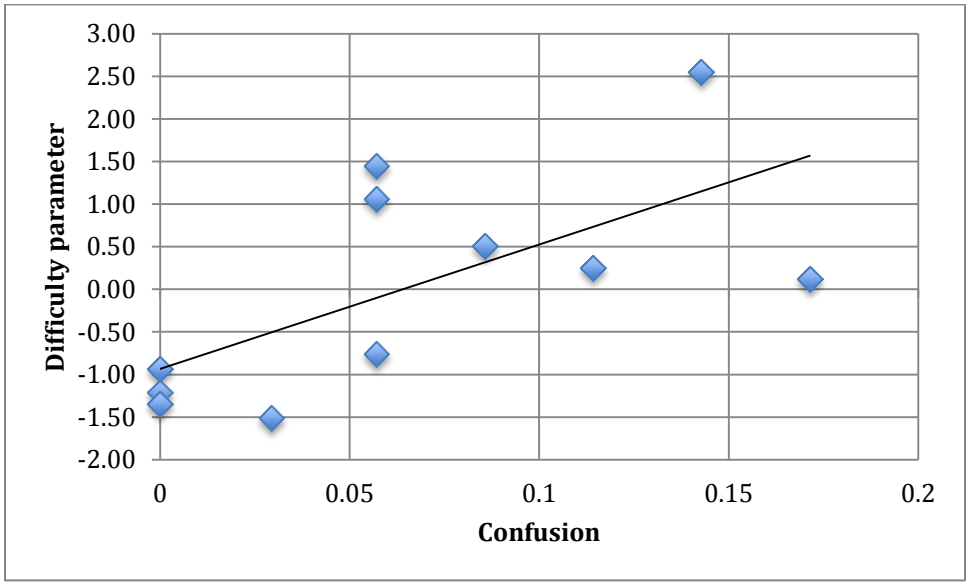

Figure 4.6 Scatterplot of difficulty parameters and verbalizations of confusion

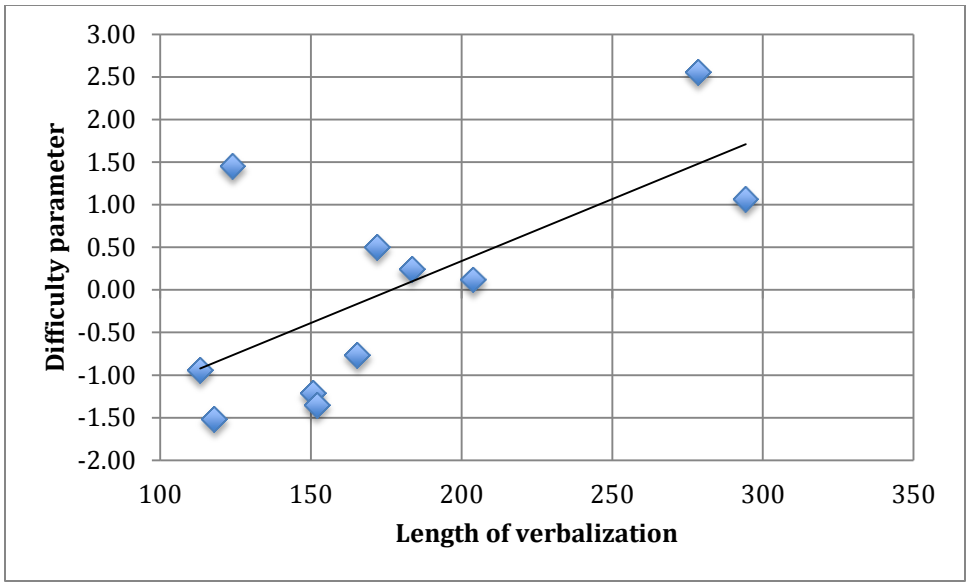

Figure 4.7 Scatterplot of difficulty parameters and length of verbalization

When p-values were correlated with the frequencies for the TAP codes for each item, nervous speech ($r= -0.769$, $p<0.01$) and confusion ($r= -0.637$, $p<0.05$) were significantly negatively correlated (see table 4.10). The correlation is expected to be negative because items that have a higher p-value (i.e. items that are less difficult) are expected to have fewer instances of nervous speech or confusion. Figures 4.8 and 4.9 display scatterplots of the relationships described above. None of the data points are considered to be outliers using Mahalanobis distance at $p<.001$.

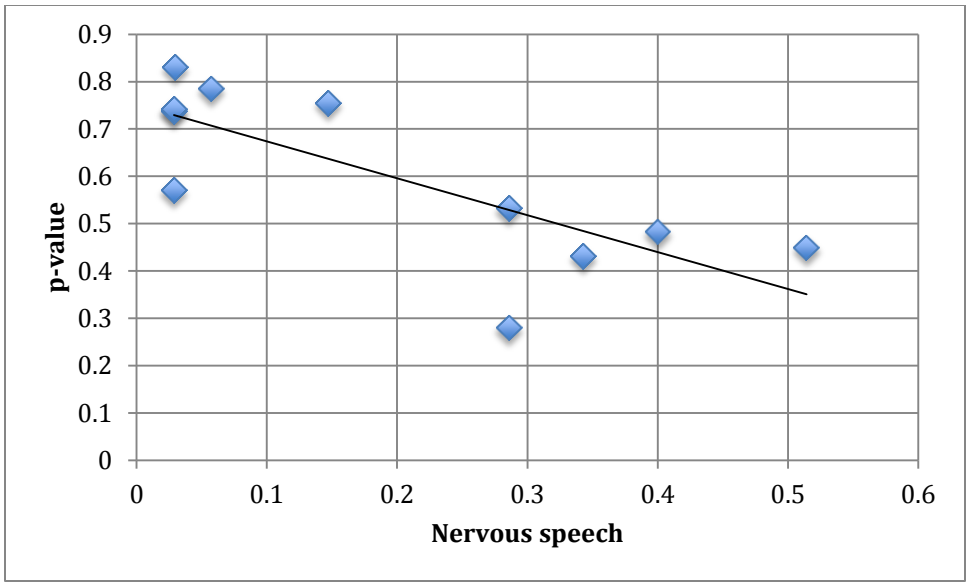

Figure 4.8 Scatterplot of p-values and nervous speech

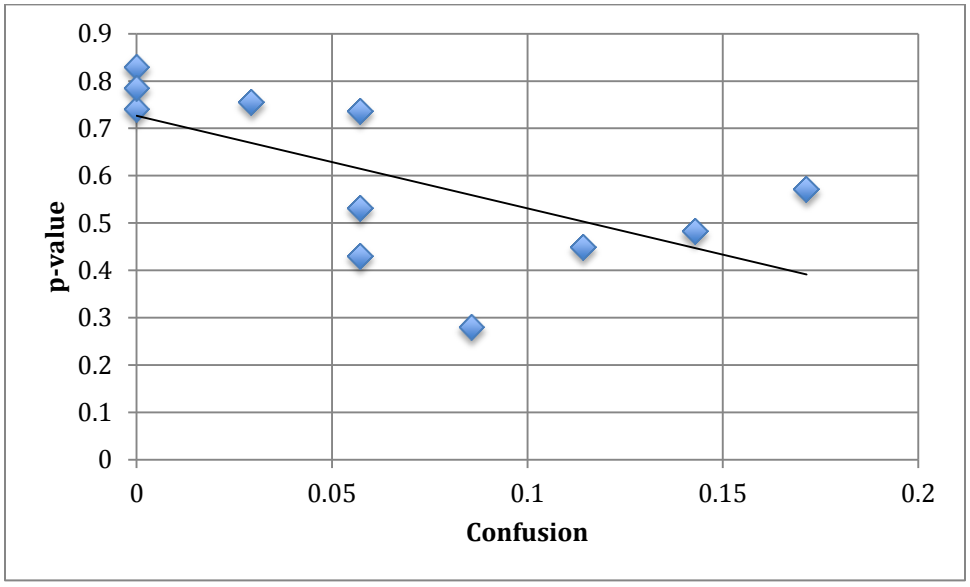

Figure 4.9 Scatterplot of p-values and confusion

4.3.2 Item Discrimination and TAPs

The next exploration of the research question involved comparing the IRT item discrimination parameters[2] with verbalizations of HT. Recall that these IRT discrimination parameters are meant to inform how well an item can "discriminate" between high and low ability students and how strongly the item taps the overall construct measured by the assessment. An item with a high discrimination parameter has a greater difference in probability of correct response (or maximum score) between high and low ability students compared to a low discrimination item.

Achievement on the HT portion of the assessment was used as an estimate of HT ability in the analysis. Based on the HT scores, the TAP sample was divided into high and low HT groups, as described in the third chapter. Differences in HT were investigated with respect to the discrimination parameter, with the expectation that the differences in evidence of HT in student verbalizations between high and low HT students would be greater for items with higher discrimination parameters. Each item was coded for different aspects of HT based on the type of HT that the item was meant to elicit. For instance, items 1 though 10 were coded for commenting on the source of the document (Source), but item 11 was not coded for this aspect of HT. Table

[2] The relationship between corrected item-total correlations and IRT discrimination parameters was first examined, just as p-values and IRT difficulty parameters had been compared in the previous section. The correlation ($r = 0.96$, $p = 0.001$) was extremely high, as expected. Because of this near-perfect correlation, it was deemed unnecessary to compare both IRT discrimination parameters and corrected item-total correlations with aspects of student verbalizations.

4.15 shows the percent of HT in student verbalizations for students who have been classified as high and low HT.

Table 4.15 Percent of students who included aspects of HT in their verbalizations

Item	Source		Perspective		Purpose		Relationship		Context	
	Low	High	Low	High	Low	High	Low	High	Low	High
1	17%	0%	58%	90%	-	-	-	-	0%	0%
2	8%	30%	33%	70%	8%	0%	-	-	25%	20%
3	75%	73%	0%	27%	0%	9%	-	-	42%	45%
4	67%	100%	92%	100%	17%	18%	100%	100%	8%	55%
5	8%	9%	33%	73%	8%	9%	8%	18%	0%	45%
6	0%	9%	0%	0%	-	-	0%	0%	8%	0%
7	25%	18%	0%	18%	17%	18%	-	-	8%	18%
8	8%	0%	17%	64%	0%	18%	0%	9%	8%	0%
9	50%	73%	25%	45%	0%	0%	0%	0%	0%	0%
10	33%	73%	50%	91%	0%	9%	50%	100%	17%	36%
11	-	-	-	-	-	-	-	-	-	-

Table 4.15 (continued) Percent of students who included aspects of HT in their verbalizations

Item	Fact		Traces		Inference		Judgment		Human nature	
	Low	High	Low	High	Low	High	Low	High	Low	High
1	17%	30%	50%	90%	17%	0%	0%	0%	-	-
2	33%	40%	42%	50%	0%	10%	-	-	-	-
3	-	-	-	-	8%	0%	0%	0%	8%	9%
4	-	-	17%	18%	0%	9%	0%	45%	-	-
5	-	-			42%	73%	0%	0%	-	-
6	-	-	25%	36%	0%	9%	-	-	-	-
7	-	-			8%	18%	0%	0%	-	-
8	-	-	0%	9%	25%	27%	8%	9%	-	-
9	-	-	-	-	0%	9%	-	-	8%	0%
10	25%	45%	25%	73%	33%	18%	0%	9%	0%	0%
11	-	-	-	-	-	-	-	-	-	-

Table 4.15 (continued) Percent of students who included aspects of HT in their verbalizations

Item	Fair		Distance		Narrative		Collective		Descendants	
	Low	High	Low	High	Low	High	Low	High	Low	High
1	-	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-	-	-	-
4	-	-	-	-	-	-	-	-	-	-
5	-	-	-	-	-	-	-	-	-	-
6	-	-	-	-	0%	0%	-	-	-	-
7	-	-	-	-	0%	0%	-	-	-	-
8	8%	0%	-	-	33%	36%	-	-	-	-
9	-	-	-	-	-	-	-	-	-	-
10	33%	45%	0%	9%	-	-	-	-	-	-
11	58%	55%	33%	64%	-	-	50%	45%	4%	4%

Table 4.15 shows that there were no individual aspects of HT that were coded for all items. However, four of the HT aspects were coded for 10 of the 11 items. These aspects were a) source, b) perspective, c) context, and d) inference. The remaining aspects of HT were only considered for eight or fewer items, and because of this, these codes were not correlated with discrimination parameters. When differences between high and low HT students' verbalizations for source, perspective, context, and inference were each correlated with discrimination parameters, none were found to be significant ($p>0.05$). These correlations are displayed in Table 4.16. However, source and perspective had moderate correlations of 0.591 and 0.439, respectively, and non-significance may be due to low sample sizes. The scatterplots for differences in each of these aspects of HT and the corresponding discrimination parameters are shown in Figures 4.10 and 4.11. Each of these scatterplots show the discrimination parameters on the y-axis and the differences between high and low HT for a particular code on the x-axis. Using Mahalanobis distance at $p<.001$, none of the data points are considered to be outliers.

Table 4.16 Correlations between discrimination parameter and aspects of HT

	Discrimination parameter
Source	0.591
Perspective	0.439
Context	0.184
Inference	-0.166

Note. Correlations were not significant ($p>0.05$)

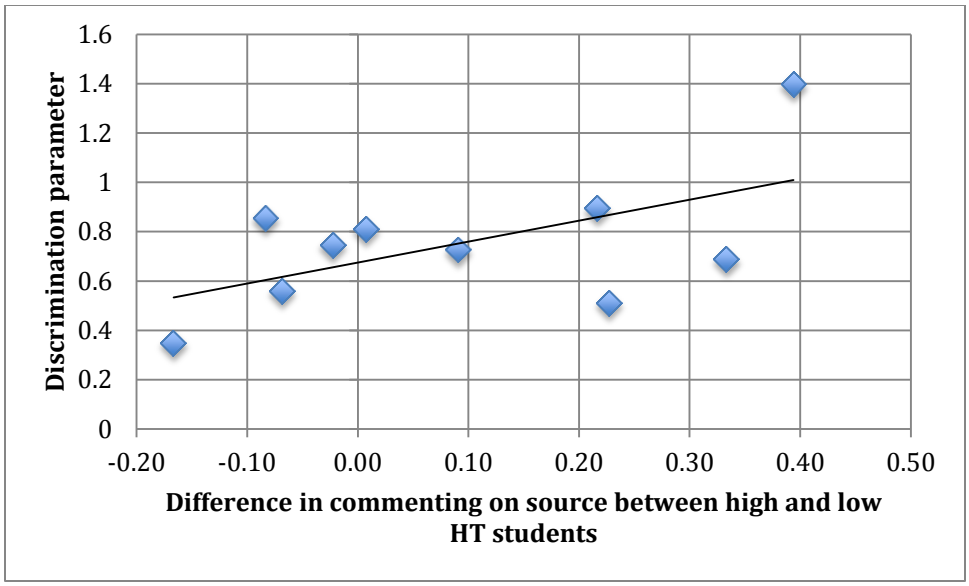

Figure 4.10 Scatterplot of differences in commenting on source and discrimination parameter

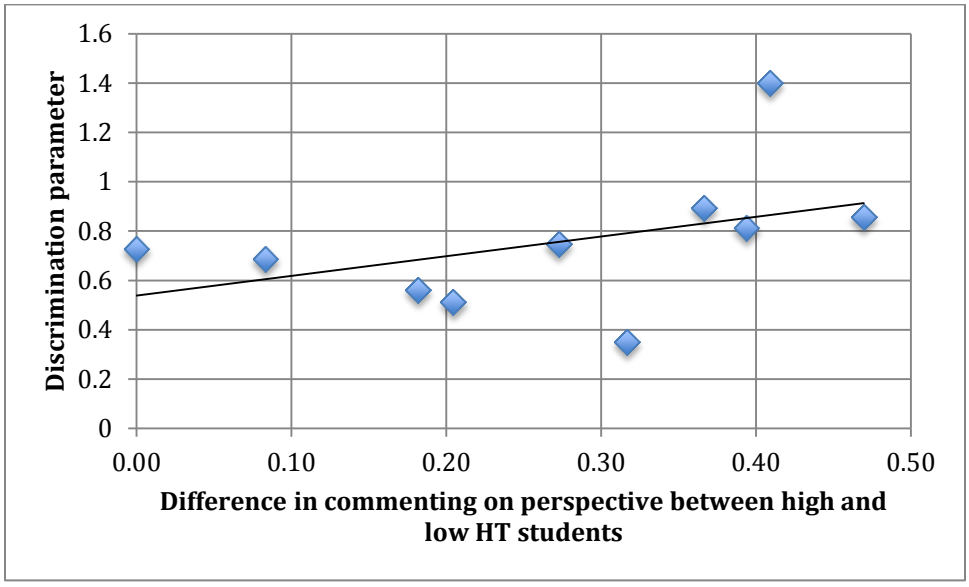

Figure 4.11 Scatterplot of differences in commenting on perspective and discrimination parameter

4.3.3 Gender Differences

Verbalizations between males and female were compared to psychometric differences between male and female examinees, including the results of the DIF analyses and p-values for

males and females. For instance, since item 4 was found to be DIF in favour of females, the TAPs of male and female students were compared to investigate if males had greater indications of difficulty in their verbalizations. Verbalizations were investigated to understand how the TAP method could provide additional evidence to the DIF results. Statistically significant differences between the two groups were not observed based on linguistic difficulty, nervous speech, guessing, not knowing, expressions of difficulty, confusion, or HT for any of the items, as discussed in section 4.2.2. However, one pattern was noticed with respect to the length of verbalizations between males and females, as displayed in Table 4.1. With the exception of the last two items, which were designed to elicit extremely long responses from students, the two greatest differences in length were the two DIF items. Though the difference in length of verbalizations between male and female examinees was not significant for any item ($p>0.05$), the item that favoured female students (item 4) had a longer average response length from females, whereas the item that favoured male students (item 7) had a longer average response length from males. In addition to the comparison between DIF items and TAPs, p-values of the items were compared between the gender groups. The differences between p-values for each gender were generally very small (less than 0.1), and a significant difference was not found between the two groups ($p>0.05$). However, item 4 stood out as having the greatest difference in p-values between males and females. The p-value for females was 0.16 higher than the p-value for males, which may correspond with the longer length of female TAPs. Table 4.17 shows the p-values for both male and female subjects.

Table 4.17 P-values for males and females

Item	Males	Females
1	0.76	0.75
2	0.82	0.84
3	0.57	0.58
4	0.36	0.52
5	0.39	0.46
6	0.74	0.73
7	0.77	0.72
8	0.28	0.28
9	0.75	0.81
10	0.44	0.52
11	0.48	0.57

4.3.4 Factor Analysis and TAPs

The factor loadings of the EFA were presented at the beginning of the chapter show that the items loaded on one factor. The factor is thought to be HT, based on the content of the assessment and how it was developed. A pattern was observed in which CR items loaded the strongest on the factor, with MC items making up the lower half of the loadings. All CR items had loadings between 0.461 and 0.730, and all MC items had loadings between 0.228 and 0.399. Items 10 and 11, which are the final two extended response items from the assessment, loaded the strongest on the factor, with loadings of 0.730 and 0.579.

The factor loadings were compared to evidence of HT from student verbalizations. This was done to investigate if factor loadings and student verbalizations provided similar information, or if student verbalizations could provide additional information about interpreting the results of the EFA. Table 4.18 shows that items 10, 11, and 4 had strong factor loadings as well as high HT (35%, 46%, and 45%, respectively) relative to the other items. However, item 5, which had the third strongest factor loading, showed a moderate mean HT of 22%. Item 8, which had the fourth strongest factor loading, had a mean HT of 15%. The remaining items each had

loadings of less that 0.4, however, the HT for these items was mixed, ranging between 9% and 26%.

Table 4.18 Item factor loadings, mean percent HT, and item type

Item	Factor loading	Percent HT	Item type
10	0.73	35%	CR
11	0.579	46%	CR
5	0.481	22%	CR
4	0.469	45%	CR
8	0.461	15%	CR
6	0.399	9%	MC
3	0.398	27%	MC
2	0.389	22%	MC
7	0.344	9%	MC
9	0.308	20%	MC
1	0.228	26%	MC

4.3.5 Item Format

The relationship between evidence obtained from psychometric information, such as EFA, p-values, and IRT difficulty parameters, and from TAPs was investigated with respect to item format. This investigation addressed the second sub-question of the primary research question, which asks how TAPs provide information about MC versus CR items, and how TAPs compare to psychometric information about item type. Furthermore, the sub-question also asks if TAPs provide information about how complex HT varies depending on item type.

4.3.5.1 Students' Experiences of Difficulty

Table 4.19 shows the mean IRT difficulty parameters and p-values for CR and MC items. This table shows that, on average, CR items were more difficult than MC items, with a lower mean p-value and a higher mean IRT difficulty parameter.

Table 4.19 Indicators of difficulty for MC and CR items

	Item type	
	MC	CR
IRT difficulty parameter	-0.944	1.162
P-value	0.737	0.435

As described in section 4.2.3, which described the validity evidence provided by TAPs about item type, CR items elicited significantly more instances of nervous speech from students. This is consistent with CR items having higher difficulty levels. However, significant differences were not found to be present between item types for other indicators of difficulty (i.e., linguistic difficulty, guessing, etc.) that were examined in this research.

4.3.5.2 Student Expressions of HT

The EFA showed that CR items loaded strongly as a group on the dimension, with MC items loading less strongly on the same dimension. The TAP findings were somewhat consistent with the EFA. Three of the four CR items that loaded the strongest on the dimension were also found to have the highest evidence of HT. However, results for the rest of the items were mixed, as discussed in section 4.3.4, which compared the results of the EFA to evidence of HT from TAPs. Student verbalizations showed that some MC items had more evidence of HT than CR items. Furthermore, there were no significant differences between MC and CR items for individual aspects of HT, as discussed in section 4.2.3, which provided validity evidence from TAPs about item type.

4.4 Student Reported Accuracy of Verbalizations

At the end of the assessment, students were asked to rate the degree to which their TAPs were similar to what they had been thinking, and then discuss possible reasons for their answers. The purpose of this portion of the research was to corroborate the accuracy of TAPs and whether differences exist in concurrent versus retrospective TAPs. Thirty-two out of the original 35 students took part in this portion of the research. The first question was concerned with how students felt that their concurrent TAPs were similar to their thoughts. Students were asked, "When you 'thought aloud' *while* you were answering the questions, how similar was what you said to what you were actually thinking?" Of the 32 students, none responded "Not at all similar" and 6% (n=2) responded "Not very similar". Most students, 72% (n=23), responded "Somewhat similar" and 22 % of students (n=7) responded that their TAPs were "Extremely similar" to what they were thinking. TAP administrators also asked, "For some questions, we asked what you had been thinking *after* you had given your answer. How similar was what you said to what you had actually been thinking?" This question differed from the previous question because it concerned how students performed retrospective TAPs, that is, verbalizations that took place after the student had answered an item. Thirty-one percent of students (n=10) answered "Extremely similar", while 19% (n=6) answered "Not very similar". Fifty percent of students (n=16) responded that their TAPs were "Somewhat similar" to what they had been thinking. Table 4.20 shows how students answered the question for concurrent and retrospective TAPs.

Table 4.20 Frequency of student responses to the similarity between their verbalizations and thought processes

Response	Concurrent	Retrospective
Extremely	7 (22%)	10 (31%)
Somewhat	23 (72%)	16 (50%)
Not very	2 (6%)	6 (19%)
Not at all	0 (0%)	0 (0%)
Total	32(100%)	32(100%)

TAP administrators were asked to explore the students' reasoning behind their responses to both questions about concurrent and retrospective verbalizations. With respect to concurrent verbalizations, the largest group of students, 53% (n=17), reported that any dissimilarity between their TAPs and thinking processes was because their thoughts were difficult to put into words. For example, one student said "I think they were pretty similar but some parts [were] hard to put in words cause in my mind it works but yeah sometimes I can't put it in words." Another student reported, "I find it fairly hard to explain my thoughts in words." Sixteen percent of students (n=5) reported that one source of dissimilarity was because their thoughts occurred faster than their verbalizations. For instance, a student noted, "it's just hard because you think so quickly and sometimes you're thinking more than one word and when you're speaking aloud you don't know which word to say and then you're just on to thinking another thing." Another student said, "usually I wouldn't talk while I'm doing a test so everything…kind of like wooshes in my mind so I can't say everything out loud right". Other students noted the effects that performing a TAP had on their thought processes. Six percent of students (n=2) reported that performing TAPs distracted them from their regular thought processes, while 9% of students (n=3) said that it helped them to think more about the problems in the assessment. Table 4.21 shows the number and percent of students who provided certain explanations about their concurrent verbalizations.

Table 4.21 Frequency of student explanations for why concurrent verbalizations were similar or dissimilar to thought processes

Explanation	Frequency (percent)
Difficult to put into words	17 (53%)
Distracting	2 (6%)
Helps with answering the question	3 (9%)
Thinking too quickly to verbalize	5 (16%)
No explanation	5 (16%)
Total	32 (100%)

Students provided possible reasons for why retrospective TAPs may be more or less similar to their thought processes. Some explanations were similar to student explanations about their concurrent verbalizations, though some explanations were specific to TAPs performed after a task was completed, and explained why those TAPS were either more or less similar to the students' thinking. In terms of why their verbalizations were dissimilar to their thoughts, 10% of students (n=3) again stated that their thoughts were difficult to put into words. Six percent of students (n=2) also said that the TAPs were difficult to do because they had never done a similar exercise before. The largest group of students, 38% (n=12), suggested that the processes described in their verbalizations changed from their actual thoughts during the task. For instance, one student said, "I changed it a little bit, because I sometimes forget, shifting my memory. I do remember what I was thinking, but it is not exactly the same wording." Another student pointed out that her verbalizations were provided in a different order than that of her thoughts, stating, "[the] chronological order of things was a little different".

Some students provided reasons for why the verbalizations provided after the task might have been more accurate than the concurrent TAPs. Six percent of students (n=2) reported that these TAPs were not distracting to their thought processes, and 10% of students (n=3) said that it was easier to say what was thought. For example, one student said, "I felt that at the end I kind of

gathered all my thoughts so I can explain it more in words instead of just…random points I guess in my head". The same number of students (n=3) reported that the TAPs helped them to think about the assessment items. Table 4.22 shows the number and percent of students who provided certain explanations about their retrospective verbalizations.

Table 4.22 Frequency of student explanations for why retrospective verbalizations were similar or dissimilar to thought processes

Explanation	Frequency (percent)
Difficult to put into words	3 (9%)
Not distracting	2 (6%)
Helps with answering the question	3 (9%)
Easier to say what was thought	3 (9%)
Changes	12 (38%)
Difficult because it was not done before	2 (6%)
No explanation	7 (22%)
Total	32 (100%)

4.4.1 Responses from Male versus Female Students

A sub-question of the investigation on accuracy of TAPS was whether or not male and female students differed in their reflections of their thinking processes. The sample of students was divided based on their reported gender and their responses to the questions about TAPs were compared. The sample of students for this research question was made up of 23 females and nine males.

When asked the first question, "When you 'thought aloud' *while* you were answering the questions, how similar was what you said to what you were actually thinking?", the proportion of female and male responses was similar (see Table 4.23). Student explanations of their responses showed that girls had more diversity in their answers (see Table 4.24), which may be related to the larger sample.

Table 4.23 Frequency of male and female responses about the similarity between their concurrent verbalizations and TAPs

Response	Male	Female
Extremely	2 (22%)	5 (22%)
Somewhat	6 (67%)	17 (74%)
Not very	1 (11%)	1 (4%)
Not at all	0 (0%)	0 (0%)
Total	9 (100%)	23(100%)

Table 4.24 Frequency of male and female explanations for why concurrent verbalizations were similar or dissimilar to thought processes

Explanation	Male	Female
Difficult to put into words	6 (67%)	11 (48%)
Distracting	-	2 (9%)
Helps with answering the question	1 (11%)	2 (9%)
Thinking too quickly to verbalize	2 (22%)	3 (13%)
No explanation	-	5 (22%)
Total	9 (100%)	23 (100%)

The second question, "For some questions, we asked what you had been thinking *after* you had given your answer. How similar was what you said to what you had actually been thinking?" showed that a greater proportion of male students were confident that their TAPs were "Extremely similar" to what they had been thinking (see Table 4.25). Finally, there appeared to be more variation in the female explanations (see Table 4.26). The greatest difference was that a larger proportion of female students who noted that their recall of their thoughts changed when they provided their TAPs after completing an item.

Table 4.25 Frequency of male and female responses to the similarity between their retrospective verbalizations and TAPs

Response	Male	Female
Extremely	5 (56%)	5 (22%)
Somewhat	3 (33%)	13 (57%)
Not very	1 (11%)	5 (22%)
Not at all	0 (0%)	0 (0%)
Total	9 (100%)	23(100%)

Table 4.26 Frequency of male and female explanations for why retrospective verbalizations were similar or dissimilar to thought processes

Explanation	Male	Female
Difficult to put into words	1 (11%)	2 (9%)
Not distracting	-	2 (9%)
Helps with answering the question	1 (11%)	2 (9%)
Easier to say what was thought	2 (22%)	1 (4%)
Changes	1 (11%)	11 (48%)
Difficult because it was not done before	-	2 (9%)
No explanation	4 (44%)	3 (13%)
Total	9 (100%)	23 (100%)

4.4.2 Responses from High versus Low HT Students

Another sub-question of the investigation of accuracy of TAPs was whether or not high versus low HT students differed in their reflections of their TAPs. The high and low HT groups that were described earlier were used for this part of the research as well. Not all of the students from these two groups took part in this portion of the research, in which they were asked to reflect on their TAPs, and so nine students made up the low HT group and 10 students made up the high HT group. In this case, both groups have very small sample sizes.

Both groups of students reacted similarly to the first question, "When you 'thought aloud' *while* you were answering the questions, how similar was what you said to what you were

actually thinking?" (see Table 4.27). There were very little differences between the two groups when students were asked to provide reasons for their answers (see Table 4.28).

Table 4.27 Frequency of high and low HT students' responses to the similarity between their concurrent verbalizations and TAPs

Response	High HT	Low HT
Extremely	1 (10%)	1 (11%)
Somewhat	8 (80%)	7 (78%)
Not very	1 (10%)	1 (11%)
Not at all	0 (0%)	0 (0%)
Total	10 (100%)	9(100%)

Table 4.28 Frequency of high and low HT students' explanations for why concurrent verbalizations were similar or dissimilar to thought processes

Explanation	High HT	Low HT
Difficult to put into words	5 (50%)	6 (67%)
Distracting	1 (10%)	-
Helps with answering the question	-	1 (11%)
Thinking too quickly to verbalize	3 (30%)	1 (11%)
No explanation	1 (10%)	1 (11%)
Total	10 (100%)	9 (100%)

When asked the second question, "For some questions, we asked what you had been thinking *after* you had given your answer. How similar was what you said to what you had actually been thinking?", students again answered similarly (see Table 4.29). In response to the follow-up question, the high and low HT groups again had similar responses (see Table 4.30).

Table 4.29 Frequency of high and low HT students' responses to the similarity between their retrospective verbalizations and TAPs

Response	High HT	Low HT
Extremely	3 (30%)	3 (33%)
Somewhat	4 (40%)	4 (44%)
Not very	3 (30%)	2 (22%)
Not at all	0 (0%)	0 (0%)
Total	10 (100%)	9(100%)

Table 4.30 Frequency of high and low HT students' explanations for why retrospective verbalizations were similar or dissimilar to thought processes

Explanation	High HT	Low HT
Difficult to put into words	1 (10%)	1 (11%)
Not distracting	1 (10%)	1 (11%)
Helps with answering the question	1 (10%)	-
Easier to say what was thought	1 (10%)	1 (11%)
Changes	3 (30%)	5 (56%)
Difficult because it was not done before	1 (10%)	-
No explanation	2 (20%)	1 (11%)
Total	10 (100%)	9 (100%)

4.5 Summary

This chapter presented the findings of each of the research questions and sub-questions. The major findings from the first research question were that the TAPs were consistent with some psychometric information, and in some cases, TAPs presented validity evidence beyond what the psychometric data could provide. The results are summarized below.

TAPs provided validity evidence about the assessment by showing how different aspects of HT occurred for individual items. For instance, TAPs showed that some items elicited more HT from students than others. TAP evidence based on gender and item type was also collected. Though there were no significant differences between gender groups, it was found that CR items

had significantly more nervous speech than MC items. TAPs also identified a number of expressions with which students experienced difficulty. That is, the verbalizations showed that a number of students struggled with terms that were important for HT in the assessment. These words were highly relevant to the subject matter of the assessment, and misunderstanding of these terms would not have been identified by other validation methods.

When verbalization codes were compared to the p-values of items, a significant negative correlation was found between p-values and nervous speech and confusion. A similar pattern was found when TAP codes were compared to the IRT difficulty parameters of items. That is, a significant correlation was found between IRT difficulty parameters and nervous speech, difficulty, confusion, and length. When IRT discrimination parameters were compared to differences in TAPs between high and low HT students, it was found that source and perspective had moderate, but non-significant correlations.

With respect to gender differences, item 4, which is a CR item that asked students to explain why an American government representative described Ukrainians differently than a Canadian religious figure, was found to favour females. Item 8, which is a CR item that asked students to determine if a Canadian politician supported the internment of Ukrainians, was found to favour males. Each of these two questions required a one-sentence answer from students. With the exception of the last two items, which were designed to elicit lengthy responses from examinees, the items with the greatest differences in length were the DIF items. When an item favoured a particular group, that group had a longer verbalization. The item that favoured females had a longer average verbalization for female students, and the item that favoured males had a longer average verbalization for male students.

EFA resulted in a one-factor model. CR items clustered more strongly on the factor, while the MC item loadings were lower. The factor loadings of each item were compared to HT evidence from TAPs. While three items had both high factor loadings and high HT, other items had mixed results. For instance, the item that had the lowest factor loading had moderate HT, and some items that had moderate factor loadings had low HT.

Further psychometric investigation of item format found that CR items were more difficult, with a lower average p-value and higher average IRT difficulty parameter. TAPs corroborated this result by showing that CR items had significantly more instances of nervous speech compared to MC items. However, when item type was compared based on HT, results showed that a number of MC items had more overall HT than some CR items, and there were no significant differences between CR and MC items for individual aspects of HT.

One major finding from the second research question, which questioned the extent to which participants believed TAPs were accurate reflections of what they had been thinking, was that students were generally confident that their TAPs were similar to their thought processes. In particular, 94% of students believed that their concurrent TAPs were "Somewhat similar" or "Extremely similar" to what they had been thinking. A greater proportion of students felt that their TAPs were "Not very similar" when done after completing an item compared to when TAPs were done concurrently. However, a greater proportion of students also said that their TAPs were "Extremely similar". In other words, some students felt that the concurrent TAPs were more similar to their thought processes, while others felt that their concurrent TAPs were less representative of what they had been thinking. Student explanations of why their TAPs were similar or not similar to their thought processes were collected. A common response was that their thoughts were difficult to put into words or, for the concurrent TAPs, their thoughts were

too quick to verbalize. Regarding the TAPs that were done after completing an item, many students mentioned that a change had occurred in the time between completing an item and verbalizing their thoughts. Male and female students were similar in their perceptions of their concurrent TAPs, however a greater proportion of males than females felt their retrospective TAPs were "Extremely similar" when done after completing an item. When high versus low HT students were compared, the two groups were very similar.

5 Discussion

5.1 Major Findings of the Research

This research investigated how TAPs contribute to the validity investigation of an assessment of complex thinking. In particular, the research examined the consistency of TAP evidence with psychometric information, in addition to how TAPs provide evidence that goes above and beyond traditional psychometric information. Furthermore, the research study investigated how students who performed TAPs perceived the accuracy of their own verbalizations. This chapter begins with a summary of the major findings from each research question and then discusses the implications of those findings. The limitations of the research are then discussed, and the chapter concludes with possible future directions for related research.

5.1.1 Summary of the Relationship between Psychometric Information and TAPs

The first research question asked how TAPs provide validity evidence that is consistent with traditional psychometric evidence, in addition to whether TAPs provide unique information about the assessment that may be overlooked by psychometric methods.

5.1.1.1 Item Difficulty and Student Verbalizations

The verbalizations identified a number of words with which the students experienced linguistic difficulty. There were significant correlations between item difficulty indicators and nervous speech, expressions of difficulty, confusion, and length of the verbalizations. The correlations with nervous speech, expressions of difficulty, and confusion indicate that students were more likely to make verbalizations that indicated their own difficulty in responding to an item when an item had a higher difficulty level. This finding demonstrates that student verbalizations reflected the difficulty and challenges that students experienced when they

responded to the assessment items. Length of verbalizations was originally included as an exploratory factor. The finding from this portion of the research suggests that students provided longer verbalizations for items that were more difficult.

5.1.1.2 Item Discrimination and Student Verbalizations

The relationship between item discrimination and student verbalizations was also examined, with the expectation that differences in HT would be greater between high and low HT students for items of higher discrimination. When the relationship between item discrimination and HT was examined, a significant relationship was not found to be present, however, a moderate correlation was found for two aspects of HT. It is possible that the non-significant relationship may be related to the small sample size of the analyses. These findings suggest that TAPs may be used to identify items that discriminate between students of different complex thinking abilities, particularly for these two aspects of HT; however, additional research is necessary to substantiate this claim.

5.1.1.3 Gender Differences

DIF analyses identified one item to be in favour of female students and one item to be in favour of male students. When TAPs were compared to DIF findings, patterns in length stood out. With the exception of the two final items of the assessment, which were extended response items and therefore required lengthier responses from the students, the two items with the greatest differences in length based on gender were also the two DIF items. More specifically, the DIF item that favoured females had a longer average response from females, whereas the item that favoured males had a longer average response from male students.

5.1.1.4 Item Format and Historical Thinking

MC and CR item types were compared using psychometric information such as EFA, p-values, b-parameters and student verbalizations. The EFA resulted in a one-factor model in which the CR items loaded very strongly together and the MC items loaded less strongly and also together. When the EFA factor loadings were compared to verbalizations of HT for individual items, it was found that MC and CR items were generally mixed for levels of HT. Three items that had high factor loadings also had high HT, and these items were CR. However, other items showed inconsistencies, such as some MC items that had moderate levels of HT, but low factor loadings, and CR items that loaded well on the factor, but had low HT. This finding contradicts the widely held notion that CR items are more effective at eliciting complex thinking from students. Taken alone, the EFA findings may have suggested that CR items are better able to capture HT, as HT is the assumed dimension from the EFA. The evidence from the verbalizations shows that this may not be the case.

In addition to the EFA results, other psychometric difference patterns were noticed between MC and CR items. Both p-values and b-parameters indicated that CR items were more difficult than MC items. The psychometric information was then compared to TAP information about CR and MC items. TAPs confirmed that students verbalized more nervous speech when attempting CR items than when they were answering MC items. So, while evidence from verbalizations contradicted the interpretation of the EFA, they corroborate the findings about difficulty and item type.

5.1.2 Summary of Perceptions of Accuracy of Student Verbalizations

The second research question was investigated to understand if students believed that their verbalizations were similar to their thought processes, and also sought to elicit possible sources

of similarity or dissimilarity from the students themselves. Overall, the vast majority of students were confident that their verbalizations were similar to their thought processes. Students also provided valuable information about why their TAPs were similar or dissimilar to their thought processes, or why retrospective TAPs were more or less similar than concurrent TAPs. Though male and female students were generally similar in their perceptions of whether their verbalizations reflected their thought processes accurately, a greater proportion of male students expressed strong confidence in their verbalizations being similar to their thought processes. Students with high and low HT scores did not vary in their perceptions of their verbalizations.

5.2 Implications of the Relationship between Psychometric Methods and TAPs

The results of this study have several implications for the use of TAPs in validity investigations, particularly for assessments of complex thinking. With respect to the first research question, the results show that TAPs can be used to provide evidence based on response processes that is consistent with psychometric information. First, the results showed that a relationship exists between verbalizations of difficulty and psychometric indicators of difficulty. In other words, both sources of validity evidence provided similar information about student difficulty with items. Nervous speech in particular stood out as an aspect of TAPs that may be particularly useful in understanding student response processes. However, based on p-values, the items used in this research were skewed to be less difficult in general. Past research by Leighton (2013) has shown that the usefulness of nervous speech may be limited for extremely difficult items. Therefore, while these results show that TAPs may be useful in other validation research for understanding student experiences of difficulty, future research should also be mindful of the extent to which these findings hold for items that are considered to be extremely challenging.

Consistencies between TAPs and psychometric information were also found for item type. Based on p-values and IRT difficulty parameters, CR items were found to be more difficult than MC items. The TAPs confirmed this finding by demonstrating that CR items had more instances of nervous speech from students. These results demonstrate that TAPs may be used in other validity investigations that examine how students respond to items based on response format, and in particular, student experiences of difficulty based on item format.

In addition to the confirmation that aspects of student verbalizations are consistent with psychometric findings, the results also show that TAPs provide validity evidence that goes above and beyond that of traditional psychometric information. In other words, the research demonstrates that TAPs provide validity evidence that is unique, and may add additional evidence to validity investigations. First, TAPs were able to identify a number of words with which students struggled, both in pronunciation and meaning. Other types of validity evidence may have been used to suggest that students experienced difficulty with items containing these words. However, student verbalizations provide unique validity evidence because this technique specifically indicates the sources of difficulty. For instance, many students had difficulty both pronouncing and understanding the words *Galician* and *Ukrainian*. These terms were a vital component of the assessment, and not understanding them would have affected a student's ability to fully comprehend the information provided in the assessment. Another word with which students struggled was the term *prejudiced.* Again, this term was central to the theme of the assessment and not knowing the meaning of this word would have affected a student's ability to engage in historical thinking.

TAPs also provided unique evidence to the validity investigation by showing that CR items did not necessarily elicit greater amounts of HT than MC items. That is, although the EFA

showed that CR items loaded stronger than MC items on the dimension thought to be HT, verbalizations from TAPs showed that CR items did not necessarily elicit more HT. A key implication of these findings is that TAPs are a useful technique in the validation of assessments because they may be used to provide additional, and in this case contrasting information, about other validation evidence. The results from the EFA alone would have suggested that CR items were better HT items. However, the HT evidence from TAPs showed that some MC items elicited more HT from students than some CR items.

5.2.1 Length of Verbalizations

Length was an aspect of verbalizations that was originally included in the research methodology as a way to compare TAPs and psychometric information. However, as indicated earlier, the variable was included in an exploratory manner because it was not clear what length of verbalizations would signify. That is, it was not known if longer verbalizations would indicate students' ease or difficulty with an item. The results from this research show that verbalization may be a significant factor when comparing TAPs to psychometric information, and therefore warrant discussion here. However, the findings about length were contradictory.

The first finding regarding length showed that students provided longer student verbalizations for items that had larger IRT difficulty parameters and smaller p-values. This finding suggests that student verbalizations are longer when student experience greater difficulty with an item. However, other findings suggest that the opposite is true. The TAPs showed that length of verbalizations from females and males was consistent with DIF findings. Primarily, the greatest differences in verbalizations between males and females, excluding the two extended response items, were the items that were found to exhibit DIF. The DIF item in favour of females was found to have longer female verbalizations and the DIF item in favour of males was found to

have longer male verbalizations. It was also found that students in the high HT group had significantly longer verbalizations than low HT students for a number of items. In each of these cases, the high HT group verbalized more than the low HT group. These findings suggest that students provide longer verbalizations when they are more likely to correctly answer an item. For instance, if a student is more confident in her ability to respond correctly to the item, it is plausible that the student will provide more detail in her verbalization.

While these results appear to be at odds with one another, one explanation is tied to the finding that CR items were, on average, more difficult. Therefore, if items with higher difficulties were CR type items, it makes sense that those items would also have longer verbalizations, and therefore a positive correlation between item difficulty and verbalization length. Indeed, an inspection of verbalization length for each item format shows that the mean verbalization length for CR items was 210 words, compared to a mean of 150 words for MC items. These findings certainly open the door for a possible future direction of research, which is to investigate how verbalization length provides information about student responses processes.

5.3 Implications of Student Reflections on TAPs

The responses from students about accuracy of their verbalizations add strength to the argument that TAPs are a viable option for collecting information about student response processes. Overall, students agreed that their TAPs were similar or very similar to their thoughts when attempting to answer assessment items. Additionally, the responses from students provide valuable insights for the research design of TAP studies. For instance, a major concern of students was that their thoughts were difficult to put into words. Another group of students expressed that they were thinking too quickly to keep up in their verbalizations. While these issues may be difficult to completely control, certain steps can be taken to diminish the effect

that they may have on the quality of the TAPs. First, practice TAPs are an important step in making sure that the student is comfortable and prepared for the type of performance that is necessary. Second, it is essential that the student not feel rushed during their session in order to fully provide verbalizations that are reflections of thinking, despite being longer than the actual thought process. One way to ensure this is to limit the number of items presented to the student. Additionally, at the beginning of the session, the administrator may want to specifically address these issues, and reassure the student to "try their best". As the work from Leighton (2013, 2011b) suggests, accuracy of TAPs is maximized when students feel at ease.

The responses from students also suggest that the type of TAPs, that is, whether TAPs are performed concurrently or retrospectively, may be better suited for some tasks than others. For instance, a large proportion of students noted that when they provided TAPs after completing an item, their account might have changed with time. Therefore, if retrospective TAPs were to be used, it would be sensible to use items that allow for a minimal time lapse between the performance and the verbalization. For instance, an item that is MC or short answer CR will allow a student to verbalize quickly after reaching an answer, whereas a long answer CR item requires a student to formulate an answer and then write it down, delaying the student's verbalization. This recommendation is consistent with research by Ericsson and Simon (1993), who advise that concurrent TAPs may be more accurate for tasks that take 7-10 minutes to complete.

5.4 Contribution of TAPs to the Validation of Assessments of Complex Thinking

This research has shown that verbalizations from TAPs can be used as validity evidence that compliments other types of evidence, including psychometric information. While evidence from TAPs should not replace other types of validity information, it contributes to the overall

validity argument that should be built about the results of an assessment by providing evidence based on response processes. That is, TAPs can, and should when appropriate, be used in conjunction with other validation evidence. TAPs contribute to the interpretive argument about the inferences of an assessment by showing that tasks engage students in the intended cognitive processes. This research makes an important contribution to the field of measurement because it shows that TAPs provide unique validity information that is specific to complex thinking. For instance, this was established when verbalizations revealed specific words, relevant to HT, with which students experienced difficulty. Verbalizations also revealed the percent of students who showed evidence of individual aspects of HT. In a validation study for an assessment of complex thinking, this type of information may be valuable because it allows the researcher to determine which items may be useful for the purposes of the assessment. In other words, this type of information can show which items elicit high levels of the desired complex thinking, and which items do not. In this study, the verbalizations also showed that some MC items had higher levels of HT than a number of CR items. Again, this type of validity evidence based on response processes is useful because it provides information about how students interact with items. In this case, it showed that item format is not necessarily a determinant for the quality of an item.

These contributions are substantial to validity research because there is a growing need for educational assessments to accurately measure the complex thinking abilities of students. As with any assessment, the interpretations of scores need to be validated. This is especially true in the case of assessments of complex thinking. TAPs provide researchers with a useful method of validating these types of assessments, in conjunction with other validation techniques.

5.5 Limitations

There are some limitations to this study that warrant discussion, though many of these limitations can be used as potential directions for improving research designs in the future. For instance, one concern about the present study is that the sample that was used for the psychometric analyses had some linguistic and cultural dissimilarities compared to the sample used for TAPs. Though the two groups have a similar proportion of students who were born in Canada, around one quarter of the TAP sample answered that their mother or father was also born in Canada, compared to over three quarters of the large-scale assessment sample. As well, the majority of students from the large-scale assessment answered that English is the most commonly spoken language in their home. In contrast, less than half of the students from the TAP sample reported that English is most commonly spoken, and a similar proportion said that Mandarin or Cantonese is the most commonly spoken language in their home. However, one question that was not asked in the student questionnaire, and one that may be considered for future studies, is the language that students consider to be their most prominently spoken language. While the TAP sample reported that Mandarin or Cantonese was most often spoken in their home, it is possible that many of these students do not consider Mandarin or Cantonese to be their primary language overall. This possibility may minimize the linguistic differences between the two samples.

Another limitation to this study is the sample size that was used for this research. Though 35 participants is an appropriate sample size for a TAP study, size became an issue when the sample was divided into subgroups. For instance, the total number of male TAP participants was 10, and the number of male students who were used to answer the second research question was nine. While this number of participants is still useful for a TAP study, it is more difficult to make

generalizations about tendencies of the group. The same is true for when high versus low HT students are compared to one another. Though smaller numbers have certainly been used for previous TAP studies, having a larger pool of students can make comparisons and generalizations more straightforward.

An additional limitation to this research is that it focused on just one kind of complex thinking. In particular, this study focused on HT, though complex thinking can occur in many other subject areas. The question arises of whether or not the findings can be generalized to other types of complex thinking For example, one finding of the research is that a number of students experienced linguistic difficulty with terms relevant to HT and the internment of Ukrainians in particular. This finding is very specific to the topic of the assessment, and therefore may not transfer easily to complex thinking more broadly. However, another finding of this research found that HT was used more in CR items compared to MC items. This finding may be more generalizable to other types of complex thinking because of the criteria used to identify HT. For instance, students had to think about the subject matter in a broader context and evaluate the information that they were given. This type of thinking is required for complex thinking in other subjects, such as mathematics (Iverson & Larson, 1996; Watson & Callingham, 2003), science (Zohar & Dori, 2003; Zohar & Tamir, 1993), and literacy (Bereiter & Scardamalia, 1987).

Focusing on Grade 11 only presents yet another limitation. Though previous TAP studies have used various age groups, the inclusion of complex thinking in this research is unique. High school students represent one particular developmental stage, and so the question remains as to whether or not these results are generalizable to younger students or adults.

A limitation of using TAPs to validate an assessment of complex thinking is that student verbalizations may not always reflect the type of thinking that the evidence is thought to

represent. That is, students may not verbalize certain aspects of HT even though they may be engaging in it. For instance, a number of students noted that their verbalizations would not keep pace with their thoughts. In some cases, students may have been engaging in complex thinking, though did not have time to articulate those processes. Conversely, it is also possible that some verbalizations were interpreted as indicating HT, despite the student not engaging in complex thinking.

With respect to student reported verification of their own TAPs, it is possible that biases that exist in students' verbalizations (Wilson, 1994) may result in biased accounts of accuracy of TAPs. One major criticism of TAPs is that, regardless of participant intentions to provide true reflections of their own thoughts, verbalizations are not accurate representations of cognitive processes. However, if the participant believes the verbalization to be accurate, then the judgment of that accuracy is also flawed. An independent account of accuracy would perhaps be necessary to confirm the participant's impressions of his own verbalization.

The scale that was used for student reported verification is an additional aspect of the research of which to be mindful. While a large proportion of students reported that their verbalizations were "somewhat" or "extremely" similar to their thoughts, it is difficult to determine the extent to which these results are artifacts of the scale itself. That is, students may shy away from extremes of the scale, and social desirability (i.e., reporting some degree of similarity between their verbalizations and thoughts) may also affect their responses.

Finally, an aspect of this research that may be considered a limitation is that the evidence from TAPs did not include the verbalizations that were provided when students were reading supporting documents. That is, in addition to verbalizing their answers, students were also asked

to think aloud while they read background information about the internment of Ukrainians, as well as primary source documents. These verbalizations were excluded because many students simply read the text of the documents without interjecting their own thoughts. The exclusion of these verbalizations may be considered a limitation because this additional set of data may have provided a more complete picture of the processes of students. However, because of the nature of TAPs, students experienced difficulty in verbalizing their thought processes while simultaneously reading aloud.

5.6 Future Directions

The overarching message that this research promotes is that TAPs are a useful method of collecting validity evidence about assessments of complex thinking, and the technique should be utilized in the future. Furthermore, the more research that is conducted using TAPs, the more information we will have regarding what TAPs can tell us about student response processes. Continuing to collect information about the strengths and limitations of TAPs will result in more balanced validity investigations moving forward.

Another direction of research is to further investigate male versus female differences in verbalizations, as well as differences between high and low HT students. Just as interviewer characteristics should be taken into account when collecting verbal reports (Leighton, 2013, 2011b; Norris, 1990), student characteristics should also be considered. While these themes were included in the present research, sample size limitations mentioned in the previous section restrict the generalizations that can be made from the findings. Future investigations would help to confirm the results, and may also reveal other findings that were unobserved in this research.

An additional future direction of research will be to investigate other types of complex thinking, both in different subject areas, as well as at different developmental stages. As

discussed earlier as a limitation, this research examined HT exclusively. Therefore, future studies that examine complex thinking in areas such as science, mathematics, and literacy would greatly contribute to what is known about using TAPs for validating assessments of complex thinking. Likewise, future research should also investigate how TAPs can be used to validate assessments of complex thinking for students in other developmental stages.

Finally, a potential area of future research is to investigate the role of length of verbalizations of students participating in TAPs. Specifically, future research should investigate if longer verbalizations reflect students' ease with an item and/or ability to correctly answer an item, as is suggested in this paper. Future research on this topic could use items that are identical in format (i.e., MC type items) so that variations in verbalization length are not influenced by that factor.

5.7 Summary

This research sought to understand if and how TAPs contribute to the validity investigation of assessments of complex thinking, above and beyond the type of evidence that is provided by traditional psychometric information. The findings from this study suggest that TAPs may be used to provide evidence that is not only consistent with psychometric information, but also provides evidence that would not have been recognized by other commonly used validation methods. Additionally, the findings from this research lend credibility to using TAPs in validity investigations. Student participants overwhelmingly agreed that their TAPs were reflective of their perceived thought processes. Furthermore, students provided explanations for their responses, which may be used in future research for guiding the accuracy of TAP methodology.

References

AERA, APA, & NCME (1999). *Standards for educational and psychological testing.* Washington DC: Authors.

Airasian, P. W., Engemann, J. F., & Gallagher, T. L. (2007). *Classroom assessment: Concepts and applications.* Toronto, ON: McGraw-Hill Ryerson.

Anderson, N. J., Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing, 8*(1), 41-66. doi: 10.1177/026553229100800104

Arbuthnot, K. (2009). The effects of stereotype threat on standardized mathematics test performance and cognitive processing. *Harvard Education Review, 79*(3), 448-472.

Ayala, C. C., Yin, Y., Schultz, S., & Shavelson, R. (2002). *On science achievement from the perspective of different types of tests: A multidimensional approach to achievement validation* (CSE Tech. Rep. No. 572). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

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

Bereiter, C., & Scardamalia, M. (1987). An attainable version of high literacy: Approaches to teaching higher-order skills in reading and writing. *Curriculum Inquiry, 17*(1), 9-30.

Bloom, B., Englehart, M. Furst, E., Hill, W., & Krathwohl, D. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain.* New York, NY: Longmans.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061-1071. doi:10.1037/0033-295X.111.4.1061

Burkett, G. (1998). Pardux (Version 1.02) [Software]. CTB/McGraw-Hill.

Calderhead, J. (1981). Stimulated recall: A method for research on teaching. *British Journal of Educational Psychology, 51*(2), 211-217.

Camerer, C.F., & Johnson, E.J. (1991). The process-performance paradox in expert judgment: How can experts know so much and predict so badly? In K.A. Ericcson & J. Smith (Eds.), *Toward a general theory of expertise* (pp. 195-217). Cambridge, England: Cambridge University Press.

Charness, N. (1981). Search in chess: Age and skill differences. *Journal of Experimental Psychology: Human Perception and Performance, 7*(2), 467.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*(3), 397-412. doi:10.1177/0013164407310130

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Core State Standards Initiative. (2010). Common Core State Standards for Mathematics. Washington, DC: National Governors Association Center for Best Practices and the Council of Chief State School Officers.

CTB/McGrawHill. (2008). *Accuracy of test scores: Why IRT models matter.* Retrieved from www.ctb.com/ctb.com/control/openFileShowAction?mediaId=18747.0

Cremeens, J., Eiser, C., & Blades, M. (2007). A qualitative investigation of school-aged children's answers to items from a generic quality of life measure. *Child: Care, Health and Development, 33*(1), 83-89.

Cromwell, L. S. (1992). Assessing critical thinking. *New Directions for Community Colleges, 77,* 37-50.

Cuban, L. (1984). Policy and research dilemmas in the teaching of reasoning: Unplanned designs. *Review of Educational Research, 54,* 655-681.

Desimone, L. M., & Le Floch, K. C. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Educational Evaluation and Policy Analysis, 26*(1), 1-22.

Ennis, R. (1987). A taxonomy of critical thinking dispositions and abilities. In J.B. Baron & R. Sternberg (Eds.), *Teaching thinking skills: Theory and practice* (pp. 9-26). New York, NY: W.H. Freeman & Co.

Ennis, R. H. (1993). Critical thinking assessment. *Theory into Practice, 32*(3), 179-186.

Ercikan, K. (2006). Developments in assessment of student learning and achievement. In P. A. Alexander and P. H. Winne (Eds.), *American Psychological Association, Division 15, Handbook of educational psychology, 2nd edition* (pp. 929-953). Mahwah, NJ: Lawrence Erlbaum Associates.

Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert review. *Educational Measurement: Issues and Practice, 29*, 24-35.

Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education, 17*, 301-321.

Ercikan, K., & Roth, W.-M. (2006). Constructing data. In C. Conrad & R. Serlin (Eds.), *SAGE Handbook for research in education: Engaging ideas and enriching inquiry* (pp. 451–475). Thousand Oaks, CA: Sage.

Ercikan, K**.,** & Seixas, P. C. (2011). Assessment of higher order thinking: The case of historical thinking. In G. Schraw and D.H. Robinson (Eds.), *Assessment of higher order thinking skills (pp. 245-261).* Charlotte, NC: Information Age Publishing.

Ercikan, K., Seixas, P. C., Lyons-Thomas, J., & Gibson, L. (April, 2012). An evidence-centered assessment design for historical thinking. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, BC.

Ericsson, K. A. (2006). Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. Hoffman. (Eds.), *Handbook of expertise and expert performance (pp. 223–241).* New York: Cambridge University Press.

Ericsson, K. A., & Simon, H. A. (1985). Protocol analysis. In T.A, van Dijk (Ed.), *Handbook of discourse analysis: Dimensions of discourse* (pp 259-268). New York, NY: Academic Press.

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

Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity, 5*(3), 178-186.

Facione, P.A. (1984). Toward a theory of critical thinking. *Liberal Education, 70*, 253-261.

Fonteyn, M. E., Kuipers, B., & Grobe, S. J. (1993). A description of think aloud method and protocol analysis. *Qualitative Health Research, 3*(4), 430-441.

Furr, R. M., & Bacharach, V. R. (2008). Psychometrics: An introduction. Thousand Oaks, CA: Sage.

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2011). Investigating the substantive aspect of construct validity for the Satisfaction with Life Scale adapted for Children: A focus on cognitive processes. *Social Indicators Research, 100*(1), 37-60.

Giancarlo-Gittens, C. (2009). Assessing critical dispositions in an ear of high stakes standardized testing. In J. Sobocan, L. Groarke, R. Johnson, & F. Ellet Jr. (Eds.). *Critical thinking education and assessment: Can higher order thinking be tested?* (pp. 17-34). London, ON: The Althouse Press.

Gierl, M.J. (1997). Comparing the cognitive representations of test developers and students on a mathematics achievement test using Bloom's taxonomy. *Journal of Educational Research, 91*, 26-32.

Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice, 25*(4), 21-35.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2*(4), 313-334.

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

Harrison, D. A., McLaughlin, M. E., & Coalter, T. M. (1996). Context, cognition, and common method variance: Psychometric and verbal protocol evidence. *Organizational Behavior and Human Decision Processes 68*(3), 246-261.

Henry, S. B., Lebreck, D. B., & Holzemer, W. L. (2007). The effect of verbalization of cognitive processes on clinical decision making. *Research in Nursing & Health, 12*(3), 187-193.

Hidalgo, M. D., & Lopez-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*(6), 903-915.

Hubley, A. M., & Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. *Journal of General Psychology, 123*(3), 207-215.

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin, 104*(1), 53.

Iversen, S.M. & Larson, C.J. (2006). Simple thinking using complex math vs. complex thinking using simple math. *Zentralblatt für Didaktik der Mathematik, 38*(3), 281-292.

Johnstone, C. J., Bottsford-Miller, N. A., & Thompson, S. J. (2006). *Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and English language learners* (Tech. Rep. No. 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319-342. doi:10.1111/j.1745-3984.2001.tb01130.x

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-64). Washington, DC: American Council on Education.

Knight, C. L. H. (1992). Teaching critical thinking in the social sciences. *New Directions for Community Colleges, 1992*(77), 63-73.

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

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

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

Lawton, C. A., Charleston, S. I., & Zieles, A. S. (1996). Individual-and gender-related differences in indoor wayfinding. *Environment and Behavior, 28*(2), 204-219.

Lee, A. M., Landin, D. K., & Carter, J. A. (1992). Student thoughts during tennis instruction. Journal of Teaching in Physical Education, 11(3), 256-267.

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* (2), 136-157. doi: 10.1080/08957347.2013.765435

Leighton, J.P. (2011a). A cognitive model of higher order thinking skills: Implications for assessment. In G. Schraw & D.H. Robinson (Eds.), *Current perspectives on cognition, learning, and instruction: Assessment of higher order thinking skills* (pp. 151-181). Charlotte, NC: Information Age Publishing.

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

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*(4), 6-15.

Leow, R. P., & Morgan-Short, K. (2004). To think aloud or not to think aloud: The issue of reactivity in SLA research methodology. *Studies in Second Language Acquisition, 26*(1), 35-57.

Levering, B. (2006). Epistemological issues in phenomenological research: How authoritative are people's accounts of their own perceptions? *Journal of Philosophy of Education, 40*(4), 451-462.

Lewis, A., & Smith, D. (1993). Defining higher order thinking. *Theory into Practice, 32*, 131-137.

Liimatainen, L., Poskiparta, M., Karhila, P., & Sjögren, A. (2001). The development of reflective learning in the context of health counselling and health promotion during nurse education. *Journal of Advanced Nursing, 34*(5), 648-658.

Linn, R. L., & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18*, 109-118.

Liu, O. L. & Wilson, M. Gender differences and similarities in PISA 2003 mathematics: A comparison between the United States and Hong Kong. *International Journal of Testing, 9*, 20-40.

Lyle, J. (2003). Stimulated recall: A report on its use in naturalistic research. *British Educational Research Journal, 29*(6), 861-878.

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences.* Stanford, CA: Stanford University Press.

Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*(303), 690-700.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Martin, J., Martin, W., Meyer, M., & Slemon, A. (1986). Empirical investigation of the cognitive mediational paradigm for research on counseling. *Journal of Counseling Psychology, 33*(2), 115.

Meade, P., & McMeniman, M. (1992). Stimulated recall—an effective methodology for examining successful teaching in science. *The Australian Educational Researcher, 19*(3), 1-18.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York, NY: Macmillan.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist, 50*(9), 741.

Miri, B., David, B. C., & Uri, Z. (2007). Purposely teaching for the promotion of higher-order thinking skills: A case of critical thinking. *Research in science education, 37*(4), 353-369.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design.* Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of Evidence-Centered design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), 6-20.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the Structure of Educational Assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-63.

Nielsen, J. (1994). Estimating the number of subjects needed for a thinking aloud test. *International Journal of Human-Computer Studies, 41*(3), 385-397.

Nielsen, J., Clemmensen, T., & Yssing, C. (2002). Getting access to what goes on in people's heads?: Reflections on the think-aloud technique. *Proceedings of the Second Nordic Conference on Human-Computer Interaction*, 101-110.

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

Norris, S. P. (1989). Can we test validly for critical thinking? *Educational Researcher, 18* (9), 21-26.

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

Paul, R. (1992). Critical thinking: What, why, and how. *New Directions for Community Colleges, 77*, 3-24.

Payne, J. (1994). Thinking aloud: Insights into information processing. *Psychological Science, 5*, 245-247.

Peck, C., & Seixas, P. (2008). Benchmarks of historical thinking: First steps. *Canadian Journal of Education, 31*(4), 1015-1038.

Penfield, R. D. (2007). DIFAS (Version 4.0) [Software]. Available from http://www.education.miami.edu/facultysites/penfield/

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement,* 17(2), 105-116.

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory & Cognition, 17*(6), 759-769.

Sands, A. (2013, March 25). Alberta's new standardized school tests will emphasize competency over content. *Global BC*. Retrieved from http://globalnews.ca/news/426363/albertas-new-standardized-school-tests-will-emphasize-competency-over-content/

Schooler, J. W., Ohlsson, S., & Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General, 122*(2), 166-183.

Seixas, P. (2010). A modest proposal for change in Canadian History Education. *International Review of History Education, 6*, 11-26.

Seixas, P. C., Ercikan, K., Gibson, L., & Lyons-Thomas, J. (April, 2012). Assessing historical thinking: Challenges and possibilities. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, BC.

Shear, B. R. & Zumbo, B. D. (April, 2012) What Counts as Evidence? An Empirical Review of Validity Studies in Educational and Psychological Measurement. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, BC.

Sireci, S. G., Patsula, L., & Hambleton, R. K. (2005). Statistical methods for identifying flaws in the test adaption process. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93–116). Mahwah, NJ: Lawrence Erlbaum Associates.

Sykes, R. C., & Yen, W. M. (2000). The scaling of mixed-item-format tests with the one-parameter and two-parameter partial credit models. *Journal of Educational Measurement, 37*(3), 221-244.

Tjeerdsma, B. L. (1997). A comparison of teacher and student perspectives of tasks and feedback. *Journal of Teaching in Physical Education, 16*(4), 388-400.

Uiterwijk, H., & Vallen, T. (2005). Linguistic sources of item bias for second generation immigrants in dutch tests. *Language Testing, 22*(2), 211-234.

van den Haak, M., De Jong, M., & Schellens, P. J. (2003). Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour and Information Technology, 22*(5), 339-351.

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

Watson, J. M., & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal, 2*(2), 3-46.

Webb, N. L. (1999, August). Alignment of science and mathematics standards and assessments in four states (NISE Research Monograph No. 18). Madison, WI: University of Wisconsin-Madison, National Institute for Science Education.

Webb, N. L. (2002, April). An analysis of the alignment between mathematics standards and assessments for three states. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Williams, A., & Davids, K. (1997). Assessing cue usage in performance contexts: A comparison between eye-movement and concurrent verbal report methods. *Behavior Research Methods, 29*(3), 364-375.

Willis, G. B., DeMaio, T. J., & Harris-Kojetin, B. (1999). Is the bandwagon headed to the methodological promised land? Evaluating the validity of cognitive interviewing techniques. In M. G. Sirken, D. J. Herrmann, S. Schechter, N. Schwarz, J. M. Tanur, & R. Tourangeau (Eds.), *Cognition and survey research* (pp. 133–153). New York, NY: Wiley.

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

Zohar, A., & Dori, Y. J. (2003). Higher order thinking skills and low-achieving students: Are they mutually exclusive? *The Journal of the Learning Sciences*, *12*(2), 145-181.

Zohar, A., & Tamir, P. (1993). Incorporating critical thinking into a regular high school biology curriculum. *School Science and Mathematics*, *93*(3), 136-140.

Zucker, S., Sassman, C., & Case, B. J. (2004). *Cognitive labs.* San Antonio, TX: Pearson.

Zumbo, B. D. (2009). Validity as Contextualized and Pragmatic Explanation, and Its Implications for Validation Practice. In Robert W. Lissitz (Ed.) *The Concept of Validity: Revisions, New Directions and Applications*, (pp. 65-82). Charlotte, NC: Information Age Publishing.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, *30*(3), 233-251.

Zwick, R. and Ercikan, K.E. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement*, *26*(1), 55-66.

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education*, *10*(4), 321-344.

Appendices

Appendix A: Assessment Tool

School Code_____

Teacher Code_____

Student Code_____

Date_____

World War I

This assessment has two sections.

In Section I, respond to the questions using your knowledge about World War I.

In Section II, there are documents as well as questions. Write your answers using the documents as well as what you learned in your social studies classes.

Multiple choice questions for both sections are numbered from 1 to 21. Mark your answers to these questions on the scantron sheets. Write your answers for the open ended questions in the test booklet itself. Good luck!

Section I

1. The immediate cause of World War I was the:
   a. march of Germany into Poland.
   b. Assassination of the Austrian Archduke Ferdinand.
   c. Sinking of the Lusitania.
   d. Moroccan Crisis.

2. The countries involved in the Triple Entente were:
   a. Germany, Austria and Italy.
   b. Germany, Austria and Russia.
   c. France, Russia and Great Britain.
   d. France, Russia and Germany.

3. The armistice ending World War I was declared on
   a. November 11, 1918
   b. June 28, 1914
   c. November 11, 1917
   d. June 28, 1919

4. The name given to the peacekeeping organization established immediately following the end of World War I was:
   a. The United Nations.
   b. The League of Nations.
   c. N.A.T.O.
   d. The Diplomatic Nations.

5. Canada was involved in World War I basically because
   a. The government of Sir Robert Borden decided to support Britain.
   b. Canada was automatically at war as a member of the British Empire.
   c. Canada signed treaties with the enemies of Germany.
   d. Canadian trade was threatened by German submarines.

6. The two different countries most involved in the arms race before World War I were
   a. Germany and France
   b. Russia and Germany
   c. Great Britain and Austria
   d. Great Britain and Germany

7. World War I and conscription are associated with which Prime Minister?
   a. W.L. Mackenzie King.
   b. R.B.Bennett.
   c. Sir Robert Borden.
   d. Sir Wilfred Laurier.

8. Which was NOT a battle fought during World War I?
   a. Vimy Ridge
   b. Somme
   c. Dieppe
   d. Ypres

Use the following events to answer question #9

1. Germany attacks Belgium
2. Archduke Ferdinand is assassinated
3. Great Britain declares war on Germany

9. What is the chronological order of these events?
   a. 1,2,3
   b. 2,1,3
   c. 2,3,1
   d. 3,1,2

10. Which of the following best describes a "war of attrition"?
    a. One side incorporates the defeated enemy's artillery.
    b. One side uses lightning warfare to rapidly gain ground.
    c. Both sides wear each other down until one is forced to give in.
    d. Both sides build up their armed forces before engaging in warfare.

11. The main result of enforcing conscription was:
    a. an increase in overseas troops.
    b. a more unified Canada.
    c. an increase in morale overseas.
    d. a major rift in English-French relations.

12. The plan which the Germans used to attack western Europe in World War One was called the
    a. Blitzkrieg Plan
    b. Belgium Plan
    c. Schlieffen Plan
    d. Hindenberg Plan

13. The War Time Elections Act of 1917:
    a. Made conscription compulsory.
    b. Extended the franchise to all men and women in the Armed Forces.
    c. Extended the franchise to female relatives of Canadians serving overseas.
    d. None of these.

14. Which of the following clauses was NOT included in the Treaty of Versailles?
    a. Germany and its allies were required to accept full responsibility for the outbreak of the war.
    b. Germany was not allowed to build any military fortifications.
    c. Germany was required to give up control of the Saar basin.
    d. Germany was not required to make any payments.

15. Canada's development as an independent nation was incurred at the end of World War I when:
    a. Canada built its own navy.
    b. Canada became more industrialized.
    c. Canada signed the Versailles Treaty as an independent country.
    d. Canada refused to send its army to other British Wars.

Section II

World War I Internment of Ukrainians
Documents and Questions

Below you will find information, documents and questions about the internment of Ukrainians during World War I. Read the background information. The questions below each document can be answered by reading the document carefully.

Background Information:

- In the 19th century the Ukrainian people did not have a nation of their own. They were divided between two powerful empires; in the west they were controlled by the **Austro-Hungarian Empire**, while the east was part of the **Russian Empire**.
- Approximately 171,000 Ukrainian people came to Canada between 1892 and 1914. The term Ukrainian was not commonly used, and immigrants arriving in Canada carried either Austrian or Russian passports.
- When Great Britain, Russia and France declared war against Germany and Austria-Hungary on August 4, 1914, Canada was automatically at war as a colony of Britain.
- The War Measures Act was passed on August 22, 1914 in order to give the government emergency powers to censor and control all publications and communications, arrest, detain or deport anyone, and take, use or control any property for the security, defence, peace, order, and welfare of Canada.
- On October 28, 1914 an Order in Council was passed requiring enemy aliens to register with authorities at different locations throughout Canada. Between October 1914 to February 24, 1920, 80,000 individuals, the majority Ukrainian, were required to report monthly to local police forces.
- In total 8,579 "enemy aliens" were sent to one of the 24 internment camps across Canada.
- Only Ukrainians originally from the Austro-Hungarian Empire and not those from the Russian Empire were sent to the camps. (Austria-Hungary was fighting Great Britain and Russia was an ally of Great Britain.)

Document 1: Attitudes towards Ukrainians 1899

An interview with Reverend Father Moris in the Calgary *Daily Herald* January 27, 1899:

As for the Galicians [Ukrainians] I have not met a single person in the whole of the North West who is sympathetic towards them. They are, from the point of view of civilization, 10 times lower than the Indians. They have not the least idea of sanitation. In their personal habits and acts, [they] resemble animals, and even in the streets of Edmonton, when they come to market, men, women, and children, would if unchecked, turn the place into a common sewer.

1. **What was Father Moris' view of Galicians:**
   a. **Most people in the Northwest are prejudiced against them.**
   b. **They are uncivilized and unclean.**
   c. **They are superior to Indians.**
   d. **Further Galician immigration should be encouraged.**

2. **This source would be useful for:**
   a. **Describing the personal habits of Galician immigrants to Canada**
   b. **Comparing how Galicians and Indians lived at this time.**
   c. **Revealing the attitudes of some Canadians towards Galician immigrants to Canada.**
   d. **Understanding how animals were treated in Edmonton.**

3. **"Father Moris' views were probably shared by many other people at the time."**
   a. **I agree with this statement because he was a church leader and his views were published in a major community newspaper.**
   b. **I agree with this statement because he had such extreme views.**
   c. **I disagree with this statement because it is just one person's view of Galician immigrants.**
   d. **I disagree with this statement because Canadians were not so prejudiced towards new immigrants.**

Document 2: American Report on the Internment of Enemy Aliens in Canada

> Under the terms of the 1907 Hague Convention neutral governments were permitted to inspect the treatment of prisoners of war being held in enemy camps. American government representative **G. Willrich** reported on prisoners of war being held in a Canadian internment camp, 29 December 1916.

The prisoners in Canadian Internment Camps came to the Dominion [of Canada] as peaceful emigrants and the great majority of them at least have been good, law abiding residents …. In other words, these men now held as prisoners … are good, sturdy, inoffensive men, able and willing to work, most of them desirous of becoming [wanting to become] Canadian citizens. .... There is no doubt in my mind, that at the present moment, the great majority of the prisoners....could safely be returned to their homes and families, and that such return would be more profitable to Canada in the end.....

4. Mr. Willrich describes the prisoners as good, law abiding residents. In one sentence explain why Mr. Willrich describes the people in the camps so differently from Father Moris (Document 1).

5. How does this source contribute to your understanding of the internment camps?

Document 3: Signed Letter from Ukrainian Newspaper Editors

This letter, signed by six Ukrainian Canadian newspaper editors was published in the *Manitoba Free Press* (Winnipeg) 17 July 1916.

The Ukrainians...of Western Canada...have found themselves heavily handicapped since the outbreak of the war by the fact of their Austrian birth which has led...the Dominion [Canadian] Government, as well as Canadian employers of labor, to unjustly class them as Austrians, and therefore enemy aliens. …They are persecuted, by thousands they are interned, [and] they are dismissed from their employment.....And why? For only one reason, that they were so unhappy as to be born into the Austrian bondage...

6. What were the six Ukrainian newspaper editors trying to explain in this letter?

a. Ukrainians are not loyal to Austria, even if they were born there.
b. Ukrainians should not be associated with Austria.
c. How Canada is treating Ukrainians unfairly.
d. All of the above.

7. Whom did the newspaper editors think was to blame for the situation they describe?

a. the Canadian government and employers
b. Austria
c. Canadian newspapers
d. none of the above: it was just part of the climate of war

Document 4: Reasons for Internment

> A speech by the Honourable C. J. Doherty, Canada's Minister of Justice, House of Commons, April 22, 1918

At the outset [beginning] of the war the Government had an option to expel the persons of enemy alien nationality….we took the position that these people....should not be allowed to leave this country. . . . Quite a number of them were interned....[more] under the inspiration of the sentiment of compassion...than because of hostility. At that time....thousands of these aliens were starving in some of our cities....we interned these people because we felt that, saying to them "You shall not leave the country" we were not entitled to say, "You shall starve within the country."

8. Did Doherty believe that the internment of Austrians was justified?

YES or NO (circle one)

Explain your answer in one sentence:

Document 5: Letter from Child to Interned Father

> Katie Domytryk, 9, to H. Domytryk, internee #1100, arrested in Edmonton, March 1916, father of four.

My dear father: We havent nothing to eat and they [government authorities] do not want to give us no wood. My mother has to go four times to get something to eat. It is better with you, because we had everything to eat. This shack is no good, my mother is going down town every day and I have to go with her and I don't go to school at winter. It is cold in that shack. We your small children kiss your hands my dear father. Goodby my dear father. Come home right away.

9. Which of the following describes how a historian investigating the internment could most likely use this source?

a. To argue that the costs of internment were justified.
b. To provide evidence of the trickiness of the Ukrainians.
c. To give a clear description of the dysfunctional families of the Ukrainians.
d. To show the impact of internment on children.

Final questions

10. Was the Canadian government justified in its policies towards Ukrainians during World War I? Using the documents and the background information, explain why or why not (one paragraph).

11. Does today's Canadian government have an obligation to make amends for internment of the Ukrainian Canadians during WWI?

Circle the sentence or sentences below that you believe answers the question.

a. There is no obligation.

b. There is an obligation and [circle one or more of the following]:

i. There should be a formal apology.
ii. The government should fund educational projects to remind all Canadians of this episode.
iii. There should be compensation to the descendants of internees for loss of wages, savings and possessions.

Provide a one-paragraph argument for the sentences that you circled.

Appendix B: Student Questionnaire

School Code: _____
Teacher Code: _____
Student Code: _____
Date: _____

World War I
Student Questionnaire

The first step in completing this questionnaire is to enter the School, Teacher and Student codes on the questionnaire (above). **Fill in the circle with the letter that corresponds to your answer.** A few questions require written responses. For these questions write your answers on the questionnaire, in the space provided.

1. Age

   Ⓐ 15 years old or younger
   Ⓑ 16 years old
   Ⓒ 17 years old
   Ⓓ 18 years old or older

2. Gender

   Ⓐ Male
   Ⓑ Female

3. Place of Birth

   Ⓐ Canada
   Ⓑ Outside of Canada

4. Mother's place of birth

   Ⓐ Canada
   Ⓑ Outside of Canada

5. Father's place of birth

   Ⓐ Canada
   Ⓑ Outside of Canada

6. Have you lived in British Columbia all of your life?

   Ⓐ Yes. I was born in British Columbia
   Ⓑ No. I moved to British Columbia before elementary school
   Ⓒ No. I moved to British Columbia after elementary school

15 October 2011

7. What is your ethnic heritage, background or national group? Select more than one or specify, if applicable.

   Ⓐ Caucasian/White
   Ⓑ Aboriginal (First Nations, Metis, or Inuit)
   Ⓒ South Asian (e.g., East Indian, Pakistani, Sri Lankan, etc.)
   Ⓓ Chinese
   Ⓔ Other (Please specify here)_____

8. Which language is most commonly used in your home?

   Ⓐ English
   Ⓑ French
   Ⓒ Mandarin or Cantonese
   Ⓓ Punjabi
   Ⓔ Other (Please specify here): _____

9. How often do people in your home talk to each other in a language other than English?

   Ⓐ Never
   Ⓑ Once in a while
   Ⓒ About half of the time
   Ⓓ All or most of the time

10. What is the highest level of schooling that **either** of your parents attended? Select only **one** response (for the parent that received the highest level of schooling).

    Ⓐ Less than high school graduation
    Ⓑ High school graduation, but no more schooling
    Ⓒ Some college or vocational training, but not a university degree
    Ⓓ University degree (for example, B.A., B.Sc.)
    Ⓔ Post-graduate professional or academic schooling

11. About how many books are there in your home?

    Ⓐ Few (0-10)
    Ⓑ Enough to fill one shelf (11-25)
    Ⓒ Enough to fill one bookcase (26-100)
    Ⓓ Enough to fill several bookcases (more than 100)

12. What mark do you usually get on social studies tests and projects?

    Ⓐ A
    Ⓑ B
    Ⓒ C+
    Ⓓ C
    Ⓔ C-
    Ⓕ I

15 October 2011

Question 13 to Question 21:
How often do the following activities take place in your history classes?

13. We listen to the teacher talk about historical events.

Ⓐ Never/Once or twice a year
Ⓑ A few times a year
Ⓒ Every month
Ⓓ Weekly
Ⓔ Almost every class

14. We are told about what was good or bad, right or wrong in history.

Ⓐ Never/Once or twice a year
Ⓑ A few times a year
Ⓒ Every month
Ⓓ Weekly
Ⓔ Almost every class

15. We discuss different explanations of what happened in the past.

Ⓐ Never/Once or twice a year
Ⓑ A few times a year
Ⓒ Every month
Ⓓ Weekly
Ⓔ Almost every class

16. We study historical sources such as letters, old documents, photographs from the past.

Ⓐ Never/Once or twice a year
Ⓑ A few times a year
Ⓒ Every month
Ⓓ Weekly
Ⓔ Almost every class

17. We watch historical videos and films.

Ⓐ Never/Once or twice a year
Ⓑ A few times a year
Ⓒ Every month
Ⓓ Weekly
Ⓔ Almost every class

15 October 2011

18. We use the textbook and/or worksheets.

Ⓐ Never/Once or twice a year
Ⓑ A few times a year
Ⓒ Every month
Ⓓ Weekly
Ⓔ Almost every class

19. We use a range of activities, e.g. role-plays, local projects or visiting museums/sites.

Ⓐ Never/Once or twice a year
Ⓑ A few times a year
Ⓒ Every month
Ⓓ Weekly
Ⓔ Almost every class

20. We retell and reinterpret history ourselves.

Ⓐ Never/Once or twice a year
Ⓑ A few times a year
Ⓒ Every month
Ⓓ Weekly
Ⓔ Almost every class

21. We use the internet and library to do historical research.

Ⓐ Never/Once or twice a year
Ⓑ A few times a year
Ⓒ Every month
Ⓓ Weekly
Ⓔ Almost every class

Question 22 to Question 32:
How important are the following goals in your history classes?

22. Learning the key facts of history

Ⓐ Not at all important
Ⓑ Somewhat important
Ⓒ Important
Ⓓ Very important

15 October 2011

23. Judging historical events in terms of ideas about human and civil rights

Ⓐ Not at all important
Ⓑ Somewhat important
Ⓒ Important
Ⓓ Very important

24. Imagining what life was like for people in the past

Ⓐ Not at all important
Ⓑ Somewhat important
Ⓒ Important
Ⓓ Very important

25. Understanding the values and decisions of people living in different situations

Ⓐ Not at all important
Ⓑ Somewhat important
Ⓒ Important
Ⓓ Very important

26. Using history to understand today's world

Ⓐ Not at all important
Ⓑ Somewhat important
Ⓒ Important
Ⓓ Very important

27. Seeing our own lives as part of a larger historical picture

Ⓐ Not at all important
Ⓑ Somewhat important
Ⓒ Important
Ⓓ Very important

28. Valuing the traditions and identity of our nation

Ⓐ Not at all important
Ⓑ Somewhat important
Ⓒ Important
Ⓓ Very important

29. Learning to value the preservation of historical sites, artifacts and old buildings

Ⓐ Not at all important
Ⓑ Somewhat important
Ⓒ Important
Ⓓ Very important

30. Learning basic democratic values

Ⓐ Not at all important
Ⓑ Somewhat important
Ⓒ Important
Ⓓ Very important

31. Learning how to judge various historical sources critically

Ⓐ Not at all important
Ⓑ Somewhat important
Ⓒ Important
Ⓓ Very important

32. Understanding diverse interpretations of history.

Ⓐ Not at all important
Ⓑ Somewhat important
Ⓒ Important
Ⓓ Very important

Please answer the following questions by writing your answers in the space provided.

33. Please list any history-related goals, not mentioned above, that you concentrate on.

34. Please write the postal code of your home address below:

___ ___ ___ ___ ___ ___

15 October 2011

Appendix C: Factor Loadings from the Two-Factor Model

Table A.1 Factor loadings from the two-factor model

Item	Factor 1 loading	Factor 2 loading	Item type
10	0.840	- 0.099	CR
11	0.618	- 0.055	CR
5	0.450	0.044	CR
8	0.429	0.055	CR
4	0.387	0.119	CR
3	0.283	0.168	MC
6	0.251	0.218	MC
7	0.235	0.161	MC
2	- 0.045	0.547	MC
1	- 0.048	0.423	MC
9	0.112	0.298	MC

Table A.2 Factor correlation matrix

Factor	1	2
1	1.00	0.540
2	0.540	1.00