Think-Aloud Protocol Procedures

TAPs involved individual administration of the assessment to students. They were asked to think aloud concurrently as they engaged in responding to each of the questions. At the beginning of the testing session the test administrator 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.

When the students were responding to questions, the test administrator noted (1) the start and end time for each question; (2) where the student was stumbling, and if the student misinterpreted the question, how the student misinterpreted the question; (3) if the student slowed down on a particular word, graphic, or part of the question; and (4) a brief version of the student's answer. If the student stopped verbalizing during a question, they were prompted to *"Remember to think aloud."* If the students' verbalizations did not include their interpretation of the question and how they came up with their response, the students were asked "*In your own words, tell me what the question asks*" and "*How did you come up with your answer to this question?*" which provide information about their understanding and thinking retrospectively. TAP administration took 48 to 118 minutes and took place in empty classrooms after the end of a school day.[2]

Sample

The TAPs and accompanying assessment were administered to a total of 35 (11 male, 24 female) students in grade 11 (10 fifteen year old students and 25 sixteen year old students). Most students (n=30) reported that they had lived in British Columbia all their lives or had moved there before elementary school. However, 34% (n=12) of students reported that Mandarin or Cantonese was the most frequently used language in their home, while 37% (n=13) indicated that English was used most commonly. With respect to previous performance in history, students were 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, two said C+, and another two said C. None

of the students reported getting lower than a C, however two students provided multiple marks. Ten of the students were part of an enriched academic program offered by the municipal school board, while the other 25 students attended a mainstream high school.

Coding of Student Verbalizations

Student verbalizations were transcribed verbatim. These transcripts were then analyzed to examine (1) whether the student understood and interpreted the tasks as intended; and (2) the extent to which the students engaged in targeted historical thinking. Both of these issues are relevant to the validity of interpreting scores as indicators of students' historical thinking. Two sets of codes were developed to interpret student verbalization in relation to these validation issues. An initial set of codes were tested with a sample of five student verbalizations and refined to make sure that the codes were clear and captured the intended meaning in verbalizations accurately. For each question, the research team defined a set of codes, which the coders used to analyze the student verbalizations. Two coders independently coded each student verbalization and recorded their codes in Excel spreadsheets prepared by the research team. After each question was completed, the coders compared codes, discussed disagreements, and reached a consensus code. The initial independent codes were recorded for examining coder agreement.

Code Set 1: Understanding of Tasks

Code Set 1 included two codes that captured understanding of the tasks. The first was the degree to which the student had a clear understanding of the question, rated as 0 to 2 for different degrees of understanding. The second was whether there were any vocabulary in the task the student did not understand, indicated by Yes or No.

Code Set 2: Historical Thinking in Student Verbalizations

For each task, we identified key historical thinking competencies and cognitive demands we expected students to engage in. These competencies and cognitive demands guided our identification of evidence of students' engagement in historical thinking in their verbalizations. For *Evidence* and *Perspective,* we identified the following types of verbalizations as evidence of or lack of historical thinking:

- *Source*: student comments on the author's identity, experience, date, or nature of the document;
- *Perspective*: student comments on the perspective of the source or its author;

- *Purpose*: student comments on the authors' purposes;
- *Comparison*: student corroborates with or contrasts to *other* documents or texts;
- *Document as Fact*: student interprets a document as fact (evidence of lack of historical thinking);
- *Traces*: student interprets sources as traces.

As evidence of *Ethical Dimension* we looked for the following in student verbalizations:

- *Fair*: student states principles of ethics or fairness (potentially, but not necessarily evidence of historical thinking);
- *Distance*: student comments on temporal distance between the time of the document and now;
- *Collective*: student builds an argument for or against the imposition of reparations (or other measures) for a historical injustice, based on considerations of collective responsibility;
- *Descendant*: student 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.

Analyzing Student Verbalizations

Coder Agreement

Inter-coder agreement Kappa (Cohen, 1960) for Code Set 1, which focused on student understanding of the questions, was very high, ranging between 80% and 100% for all codes across the 11 tasks, except for Tasks 2 and 8 for coding Understanding of the Question (UN), which were 68% and 54% respectively. Code Set 2, which required coders to make judgments about evidence of students' historical thinking, was highly challenging. Inter-coder agreement for Code Set 2 was lower than that for Code Set 1 but tended to be moderate for most of the tasks, ranging between 60% and 70%, though for some tasks it was as high as 100%, and in a handful of cases around the 30% to 40% range. These tended to be the codes that required greater interpretation of verbalizations rather than direct observations of evidence of historical thinking.

Understanding of Tasks

The student verbalizations indicated that the great majority of the students understood what the questions were asking them to do or respond to. On all tasks, except for Tasks 2 and 8, student verbalizations indicated full understanding of questions for over 70% of the students. On Task 2, 68% and on Task 8, 51% of students' verbalizations indicated full understanding of the questions. Further

examination indicated that poor understanding of Tasks 2 and 8 was not caused by confusion about the wording in the question. Instead it was caused by either a lack of knowledge about how primary sources are used in history, or confusion about whether the question was asking about the author's perspective versus the student's own perspective.

Evidence of Historical Thinking

Once the verbalizations are coded, using these codes as evidence of historical thinking requires a systematic analysis of the codes. There were three steps in this process. The first step was to determine whether student verbalizations included codes identified as evidence of either *Evidence* and *Perspective* or *Ethical Dimension*. This information is valuable in understanding what types of evidence verbalizations included. Since each task may include evidence of more than one code, for example by commenting on the perspective of the source or its author (*Perspective*) as well as interpreting sources as traces (*Traces*), evidence of both of these would provide supporting validity evidence that the task measures historical thinking. Therefore, as part of a validity investigation, the second step is to determine to what extent *any* of the relevant codes were included in the verbalizations. For example, if *Perspective* and *Traces* were the relevant codes, the second step would determine what percentage of the students included evidence of either or both of these aspects of historical thinking. This additional level of summary would therefore reflect the students who included evidence of *Perspective*, evidence of *Traces*, and those that included both aspects of historical thinking.

In order for particular verbalizations to be interpreted as evidence of historical thinking, such verbalizations should be observed for students who have higher historical thinking scores, and they should not be observed for those students who did not score well on these tasks. The consistency of inferences from verbalizations and student responses to tasks is necessary for meaningful interpretation of scores. To verify this relationship between verbalizations and scores, the third step involved comparing historical thinking scores of students who included the relevant codes of historical thinking in their verbalizations and those who did not. Each of these three steps in our research are summarized below.

Step 1: Evidence of Historical Thinking in Verbalizations Separately by Code

Evidence and Perspective

In our research, evidence of historical thinking demanded by each task was first summarized by the percentage of students who included the relevant verbalizations in their TAPs. Table 13.1 summarizes evidence of historical thinking in

student verbalizations for each code in each task. Greater percentages for each code indicate that higher proportions of students included these codes in their verbalizations and therefore constitute stronger evidence of historical thinking demanded by these tasks compared to the other tasks.

Students were expected to demonstrate *Evidence* and *Perspective* competencies on Tasks 1 to 9. There was a great degree of variability of evidence across the nine tasks. Evidence of sourcing varied from question to question, with 6% to 89% of students commenting on the author's identity, experience, date, or nature of the document (*Source*) in their verbalizations. On most of the *Evidence* and *Perspective* tasks, students commented on the perspective of the source or its author (*Perspective*) with 43% to 91% students making such comments in their verbalization of these tasks, except for three of the tasks in which only a small proportion of students made such comments. On one question, 29% of the students commented on historical worldviews or contexts of the events and information presented to them in the documents (*Context*). Only small proportions of students commented on authors' purposes (*Purpose:* 2% to 17%).

Students were expected to corroborate with or contrast documents on only three of the tasks (*Compare*). On two of these tasks, the great majority of students (100% and 74%) corroborated and contrasted documents, and on one task, only 20% verbalized corroboration or contrasting.

For evidence of historical thinking, students were expected to interpret sources as traces (*Traces*) and not read documents as fact (*Document as Fact*). Larger proportions (31% to 71%) of students provided evidence that they were aware of sources as traces, than students who read documents as facts (14% to 44%) across the nine tasks.

Ethical Dimension

In responding to questions about ethical judgment (Tasks 10 and 11), students stating general principles of ethics or fairness (*Fair*) to justify their responses could not *prima facie* be considered evidence of historical thinking or lack thereof. In question 10, if students used such statements while remarking on the historical context, or in question 11, if they used such statements qualified by recognition of the temporal distance between now and World War I, then they were interpreted as providing evidence of historical thinking. If these two qualifiers were absent in their responses to the two questions, respectively, then general principles of fairness were not considered to be evidence of historical thinking. In the two questions assessing ethical judgment, 37% and 49% demonstrated such reasoning. As evidence of understanding the ethical dimension of historical interpretations, students were expected to comment on the temporal distance between now and then (*Distance*). While more than half of the students (54%) made such comments on one of the questions, only a small proportion (6%) verbalized such comments when responding to the other question. In responding

TABLE 13.1 Evidence of historical thinking in student verbalizations by code

Task	*Codes*	*Percentage expressed in verbalization*	*Task*	*Codes*	*Percentage expressed in verbalization*
1 (MC)	Perspective*	65	10 (CR)	Comparison**	74
	Traces*	71		Context*	29
2 (MC)	Purpose*	3		Document as Fact	34
	Perspective*	44		Traces**	57
	Document as Fact*	44		Fair	37
	Traces*	32		Distance**	6
3 (MC)	Source**	77	11(CR)	Fair*	49
4 (CR)	Source**	89		Distance*	54
	Perspective*	91		Collective*	37
	Purpose*	17		Descendants**	46
	Comparison	100			
5 (CR)	Source*	6			
	Perspective*	43			
	Purpose*	9			
	Comparison**	20			
6 (MC)	Source*	20			
	Perspective*	3			
	Document as Fact*	29			
	Traces*	31			
7 (MC)	Source*	26			
	Perspective*	9			
	Purpose*	14			
	Document as Fact*	14			
8 (CR)	Perspective**	4			
	Purpose*	11			
9 (MC)	Source*	66			
	Perspective*	43			
	Purpose*	2			

*indicates that the scores were higher for students who included evidence of historical thinking in their verbalizations; **indicates statistically significant mean differences at alpha = 0.05 level for two student groups who included evidence of historical thinking and those who did not.

to the last question on reparations for Ukrainian internment in Canada, students were expected to build an argument for or against the imposition of reparations (or other measures) for a historical injustice, based on considerations of (1) collective responsibility (*Collective*); and (2) benefits and deficits to respective present-day descendants (*Descendants*). Fewer than half of the students (37% *Collective*, 46% *Descendants*) made arguments using these considerations.

Step 2: Evidence of Historical Thinking in Verbalizations Combined Across Codes

The previous section summarized evidence of historical thinking separately by code for each task. In this section, such evidence is combined across codes for each task resulting in the percentage of students who included at least one relevant aspect of historical thinking for each task (though it could also consist of students whose verbalizations included multiple relevant aspects of historical thinking). The percentage of students who provided evidence of historical thinking varied between 32% (for Task 7) to 100% (for Task 4). The Task 7 with the lowest evidence of historical thinking asked students to choose one of four options that answered "**Whom did the newspaper editors think was to blame for the situation they describe?**" based on a brief excerpt from a letter signed by six Ukrainian Canadian newspaper editors. On a closer look, answering this item correctly required students to read and understand what was presented in the excerpt without necessarily exercising historical thinking. The task with the highest evidence of historical thinking, Task 4, asked students to provide an explanation for differences in perspectives between an American government official and a religious leader presented in two separate documents: "**Mr. Willrich describes the Ukrainian prisoners as good, law abiding residents. In one sentence explain why Mr. Willrich describes Ukrainians so differently from Father Moris.**" In this task, students were explicitly required to compare perspectives in two documents and, not surprisingly, all students included comparisons of perspectives in their verbalizations.

Tasks 4, 5, 8, 10, and 11 are CR items. Even though two of these five tasks (4 and 11) had the highest percentage of students demonstrating evidence of historical thinking, some of the MC items, e.g., tasks 1, 2, 3, and 9, also had strong evidence of historical thinking and were stronger than three of the CR tasks (5, 8, and 10) (See Figure 13.2). Based on this step of the analyses, there was not consistently stronger evidence of historical thinking on CR items.

Step 3: Correspondence Between Evidence of Verbalization and Performance

If the verbalizations indicated evidence of historical thinking, then students who demonstrated historical thinking in their verbalizations would be expected to have

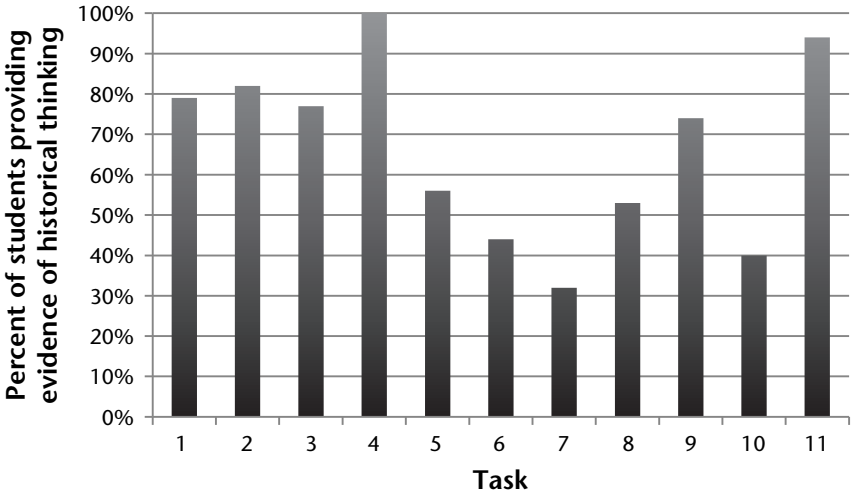

FIGURE 13.2 Percentage of students providing evidence of historical thinking in their verbalizations for each of the eleven tasks

higher scores on their written responses to those tasks. In Table 13.1, '*' indicates that the scores were higher for students who included evidence of historical thinking in their verbalizations and '**' indicates that the differences in score means were between high and low scoring students statistically significant at alpha = 0.05 level. On 36 codes, across 11 tasks, there were statistically significantly different score differences on six codes. In 25 of the comparisons, the differences were in the direction supporting historical thinking but were not statistically significant. This was not surprising given the low sample size of 35. In all, there were 2 codes (*Comparison* on task 4 and *Fair* on Task 10) for which either there were no score differences between students who provided evidence of historical thinking in their verbalizations and those who did not or they were not in the expected direction. Corroborating or contrasting (*Comparison*) on Task 4 was included in all the student verbalizations because the question specifically asked them to compare information presented in two documents. Therefore, no relationship between this evidence of historical thinking and historical thinking scores could be established because everyone, whether they were employing good or poor levels of historical thinking, included it in their verbalizations. Stating general principles of fairness (*Fair*) on Task 10 could be considered as evidence of lack of historical thinking. Task 10 asked students to discuss whether the Canadian government was justified in their policies toward Ukrainians. If students discussed contrasting perspectives in the documents and accurately explained how each is relevant to the justifiability or unjustifiability of the policies, then they would have obtained the maximum score of 3 even if their verbalizations indicated they referred to broad fairness principles. In other words, verbalizations classified as *Fair* was not clear evidence of lack of historical thinking.

Based on the analyses in this step, there was stronger evidence of historical thinking from student verbalizations for CR tasks than for MC tasks. While on all of the five CR tasks at least one code had a statistically significant association with scores based on students' written responses, only one MC task had such a relationship.

Implications for Validating Assessments of Historical Thinking

Data from TAPs provided clear cognitive evidence that the tasks in the assessment engaged students in historical thinking. Without such data, it would not have been possible to demonstrate whether the tasks measured the intended constructs. The first step of the analyses of verbalizations determined what types of historical evidence each task elucidated. This is a necessary step to understand the constructs captured by the tasks. The second step of the analyses provided information about which tasks required historical thinking from students more consistently. Such information is useful in the assessment design stage for revising or selecting tasks so that tasks with strong and consistent historical thinking requirements can be included in the assessment. In the third step, examining the relationship between evidence of historical thinking in student verbalizations and historical thinking scores demonstrated a consistent pattern for the great majority of codes across the tasks (except for three). Even when relatively small proportions of students expressed particular evidence of historical thinking in some questions, these were associated with higher scores on these tasks. On three tasks, these differences were statistically significant. Overall, the three steps of analyses provided complementary information about what the tasks were measuring.

The TAP methodology has several limitations that one needs to be aware of in using it in validation research. The first, as noted by Kaliski et al. (this volume), is that due to the labor-intensive nature of the procedure, the sample size that can be included in this type of research is limited. The small sample size also limits the strengths of inferences that can be made. For example, statistical significance may not be obtained even when there are strong systematic relationships, and moderate or weak associations may not be observed. Secondly, there is not a one to one relationship between student verbalization and evidence of competency. There are many reasons why students may or may not verbalize, including their willingness and ability to communicate their thinking, their metacognitive ability to be aware of their thinking, and the extent to which the task lends itself to the type of verbalization needed, among others (Leighton, 2011). Another issue to consider is that the tasks with the highest percentage of students including evidence of historical thinking cannot be considered as the best tasks for measuring historical thinking. In our research, Task 4 had 100% of students including comparing and contrasting perspectives in their verbalizations. This item can be considered as capturing the most basic levels of historical thinking students demonstrated by following specific instructions in

the task. Other more difficult tasks which are targeted to capture higher levels of historical thinking may not include evidence of historical thinking in verbalizations by students whose historical thinking levels may not be sufficiently high to manage the task. The third step in our analyses, which connects verbalization evidence with performance, provides better evaluation of the degree to which verbalizations were good indicators of historical thinking. Based on the findings from our research, TAPs provide necessary validity evidence for assessments of historical thinking. Without such evidence, any assessment of historical thinking will have a major gap in supporting claims about what the assessment is truly measuring.

Notes

1 www.historicalthinking.ca
2 This time includes administration of a short test with 15 multiple-choice factual knowledge questions on World War I.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

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.

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 reviews. *Educational Measurement: Issues and Practice, 29*(2), 24–35.

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

Ercikan, K., Seixas, P., Lyons-Thomas, J., & Gibson, L. (2012, March). *Designing and validating an assessment of historical thinking using evidence centered assessment design.* Paper presented at the annual meeting of the American Educational Research Association, Vancouver, BC.

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

Ferrara, S., & Chen, J. (2011, April). *Evidence for the accuracy of item response demand coding categories in think aloud verbal transcripts.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Ferrara, S., & Duncan, T. (2011). Comparing science achievement constructs: Targeted and achieved. *Education Forum, 75*, 143–156.

Ferrara, S., Duncan, T. G., Freed, R., Velez-Paschke, A., McGivern, J., Mushlin, S., Mattessich, A., Rogers, A., & Westphalen, K. (2004, April). *Examining test score validity by examining item construct validity: Preliminary analysis of evidence of the alignment of targeted and observed content, skills, and cognitive processes in a middle school science assessment.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Ferrara, S., Svetina, D., Skucha, S., & Davidson, A. H. (2011). Test development with performance standards and achievement growth in mind. *Educational Measurement: Issues and Practice, 30*(4), 3–15.

Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement, 30*(5), 394–411.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice, 18*(4), 5–9.

Huff, K., & Ferrara, S. (2010, June). *Frameworks for considering item response demands and item difficulty.* Presentation at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Detroit, MI.

Kaliski, P., France, M., Huff, K., & Thurber, A. (2010, April). *Using think aloud interviews in evidence-centered assessment design for the AP World History exam.* Paper presented at the annual conference of the American Educational Research Association, Denver, CO.

Leighton, J. P. (2011). *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., & Gierl, M. J. (2007). Verbal reports as data for Cognitive Diagnostic Assessment. In J. P. Leighton and M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 146–172). New York, NY: Cambridge University Press.

Magone, M. E., Cai, J., Silver, E. A., & Wang, N. (1994). Validating the cognitive complexity and content validity of a mathematics performance assessment. *International Journal of Educational Research, 21*, 317–340.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5–11.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–63.

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

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

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

Pottier P., Hardouin, J.-B., Hodges, B. D., Pistorius, M.-A., Connault, J., Durant, C., . . . Planchon, B. (2010). Exploring how students think: a new method combining think-aloud and concept mapping protocols. *Medical Education, 44*(9), 926–935.

Sato, E. (2011, March). *Cognitive interviews of English language learners and students with disabilities and features contributing to item difficulty: Implications for item and test design.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Seixas, P. (2009). A modest proposal for change in Canadian history education. *Teaching History, 137*, 26–31.

14

MEASURING UP?

Multiple-Choice Questions

Gabriel A. Reich

Models of Achievement in History

In history education, there are several competing theoretical models of disciplinary achievement. Within the scholarly community, a loose consensus exists about some key disciplinary concepts that can enhance students' abilities to achieve a more nuanced understanding of history (cf. Lee, 2005; Lévesque, 2008; Seixas, 1996; Wineburg, 2001). Although grounded in empirical research, this theory has only a tangential relationship with another theoretical model of disciplinary achievement, official content standards.

Official content standards are produced by education bureaucracies. They may be influenced by the history education community, but they are developed in a different institutional context, with different imperatives, mandates, and political considerations (Broadfoot, 1996; Wineburg, 1991). Far from being merely technocratic, defining content standards is a political process, one that must contend with public anxiety about the transmission of heritage and culture to the next generation (VanSledright, 2008). As institutions that are accountable to the public, education bureaucracies tend to be careful not to violate the expectations of citizens, especially in the case of history (Nash, Crabtree, & Dunn, 1997; Zimmerman, 2002).

The research reported in this chapter took place in New York State. At the time data was collected for this study, The New York State Education Department (NYSED) had published two key documents that served as the guideposts for what students were expected to know and do upon completion of the global history and geography course: the "Core Curriculum" (NYSED, 1999a) and the "Standards and Performance Indicators" (NYSED, 1999b). The "Core Curriculum" (NYSED, 1999a) is a list of content that teachers are supposed to cover in the first two years of high school. The historical information that appears on this

list varies from factual material, such as "the Marshall Plan" or the "Truman doctrine" (NYSED, 1999a, p. 113), to concepts, such as "surrogate superpower rivalries" (p. 113), and terms that denote larger narratives, such as "emergence of the superpowers" (p. 113). The "Standards and Performance Indicators" (NYSED, 1999b) present a model of achievement in history that consists of the conceptual understandings and historical thinking skills that history education should foster.

The primary purpose of state-sponsored examinations is to collect evidence that can be used to inform an argument about whether or not learning standards have been mastered by a population of students at a particular point in their education careers. To observe whether or not test-takers have met a set of standards, a task must be designed that elicits a performance that can be reasonably interpreted as an indication that the material was indeed learned (Pellegrino et al., 2001). Multiple-choice tests produce data collected under standardized conditions that can be used to make inferences about large populations of students. Stakeholders interpret test performances and use test scores to inform judgments about the effectiveness of teaching and learning (Linn, 2003; Pellegrino et al., 2001). However, the multiple-choice format includes no evidence of test-taker reasoning. Thus, a teacher may observe that students performed poorly on an exam, but the nature of the task occluded the possibility of more nuanced interpretations of what misunderstandings, for example, persist.

In New York, the state defines what it believes it is measuring when testing with multiple-choice questions in a document called the "Test Sampler Draft" (NYSED, 1999c). In it, the test developers explain that the multiple-choice questions sample from the list of content in the Core Curriculum (NYSED, 1999a). They explain further that

> the multiple-choice items are designed to assess students' understanding of content and their ability to apply this content understanding to the interpretation and analysis of graphs, cartoons, maps, charts, and diagrams. (NYSED, 1999c, p. 1)

The report also says that achievement of the more conceptual and skills-based standards (NYSED, 1999b) are measured by the thematic and document-based essays on the exam. The multiple-choice section of the exam is worth 55% of the final scaled score, and the two essays are worth 45% of the final scaled score (NYSED, 1999c).

The Study

The study described below was designed to collect evidence that informs an argument about the kinds of performances that multiple-choice history questions elicit. Scholars with an interest in test-score interpretation, or validity, have called for such research (Black, 2000; Hamilton, Nussbaum, & Snow, 1997;