Using Think Aloud Interviews in Evidence-Centered Assessment Design for the AP World History Exam

Pamela Kaliski, The College Board

Megan France, James Madison University

Kristen Huff, Regents Research Fund

Allison Thurber, The College Board

CollegeBoard inspiring minds

Importance of cognitive models

- Cognitive models of learning can guide assessment design

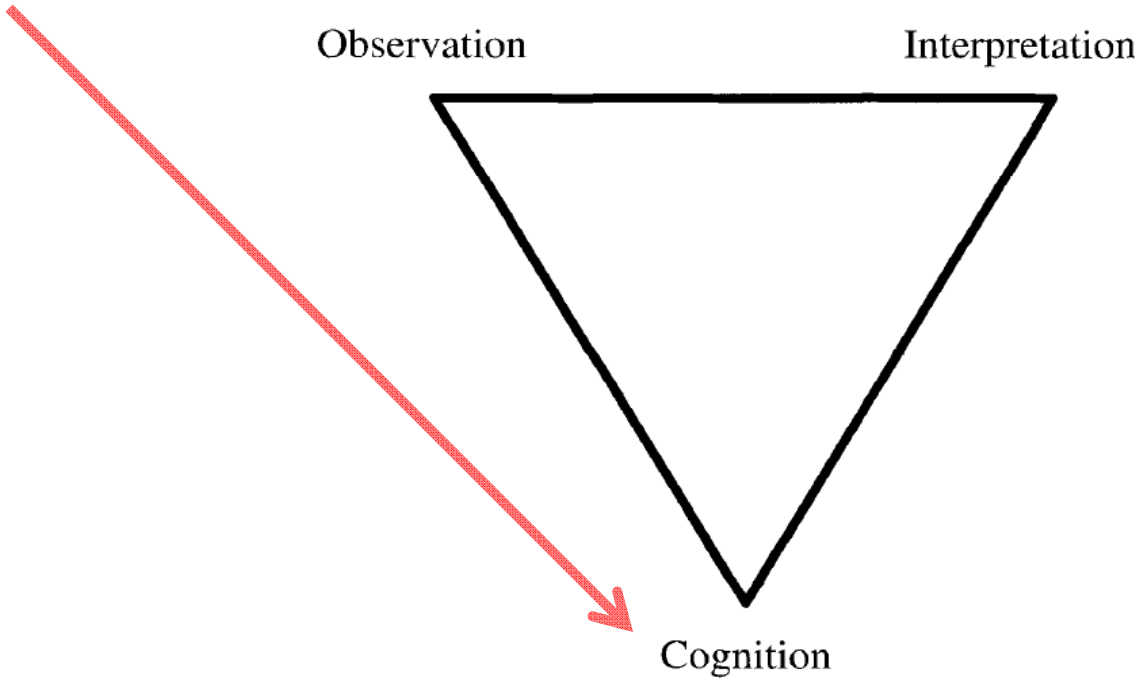


FIGURE 1. The assessment triangle. Adapted from National Research Council, 2001, p. 44.

(Leighton, 2004; NRC, 2001)

Importance of think aloud interviews—validity evidence!

- Think aloud interviews gather an important form of validity evidence that can be gathered during test development, before scores are actually obtained.
  - Are items actually eliciting the intended skills?

"If test items are being systematically misunderstood, this would mean that (a) the assessment is eliciting content understandings and processes other than what was intended, or (b) the inferences drawn from the scores are inaccurate, or both (Leighton, 2004, p.8)."

CollegeBoard
inspiring minds

Using think aloud interviews to inform item development for AP World History

- The AP program
  - ~34 courses and exams for high school students—provide college credit and placement; score of 1, 2, 3, 4, 5
- Current World History exam (70 MC items, 3FR)
  - Factual recall items are included on MC section
- New, ECD based AP World History exam
  - Several ECD phases—currently in task model development phase
    - Use claims & evidence, difficulty drivers, for item writing
  - Designed to assess a student's ability to apply historical thinking skills to historical content—9 historical thinking skills in domain
    - Are intended skills actually being measured?

(Ewing et al., 2010; Hendricskon et al., 2010; Huff, et al., 2010)

CollegeBoard inspiring minds

9 historical thinking skills

- Historical Argumentation
- Use of Evidence
- Historical Interpretation
- Historical Causation
- Comparison
- Contextualization
- Continuity and Change Over Time
- Periodization
- Synthesis

CollegeBoard
inspiring minds

Purpose of Study

1. Do the verbal reports for the new ECD-based AP World History items elicit historical reasoning skills, compared to the AP World History items on previous exams?
2. Do these piloted AP World History items elicit evidence of the intended historical thinking skills? That is, what degree of alignment exists between intended HTS and observed HTS?
3. What item features contribute to the perceived difficulty of these items? That is, what difficulty drivers are present for these items?

CollegeBoard
inspiring minds

Participants and Procedures

- 17 students who completed AP WH in 2009-2010
  - Received $50 gift card
- 15 MC items administered
  - 2 old items
  - 13 new items from 12 task models
- Researchers followed a script; practiced thinking aloud for students, then allowed practice items for students before actually starting
- Two parts for each item: Concurrent think aloud, Retrospective think aloud
- 6 verbal reports have been coded so far
  - Two 3's, two 4's, two 5's
  - Three males and three females

CollegeBoard
inspiring minds

Coding Framework

GROUP A: Cognitive Processing		
CODE	DEFINITION	EXAMPLE
Factual Recall	There is not a specific historical thinking skill required to answer this correctly, just recall. The student reads through the question once and selects the answer; no other response strategy (e.g., process of elimination, guessing) is needed. Student "knows" the answer	"No A doesn't make sense. It's not B. C is not the answer either. Yes, answer D looks right; the answer is D."
Historical Thinking Skill (HTS)	One or more of the nine historical thinking skills is utilized when interacting with the task.	See HTS codes below for examples of each HTS.
Guessing	The student does not have the knowledge/skills necessary to answer the question and just guesses. Student clearly does not know the answer	"I really have no idea. It could be A or C. It could be D too. So I'll just pick C. It looks like it could be right."
Process of Elimination	Student eliminates at least one response option to help arrive at their answer choice	
Background Characteristic	Student knows answer due to personal background (e.g., religion)	The student knows something beyond AP World History that helps lead him/her to the answer (e.g., "I'm Muslim and I know that would be in a mosque.").

Coding Framework, cont.

GROUP B: Difficulty Drivers	
CODE	DEFINITION
Length of item	Actual length of item is indicated as a reason for an item being difficult.
Characteristics of Stimulus Material	The presence of stimulus (e.g., quote, graph) is indicated as a reason for an item being difficult. Some stimulus materials are more challenging than others. Also, the length of a stimulus material affects difficulty.
Degree of Familiarity	Less familiar content is more complex than more familiar content; students have not had the opportunity to learn the content, making an item more difficult.
Quality of Distracters	Student notes the quality of distracters, they may indicate that some distracters were easy to eliminate or that a lot of the distracters seem like plausible options.
Challenging Vocabulary	Student states the vocabulary made the item difficult
Scaffolding	Student states the item was easier due to some specific detail (e.g., presence of a date)

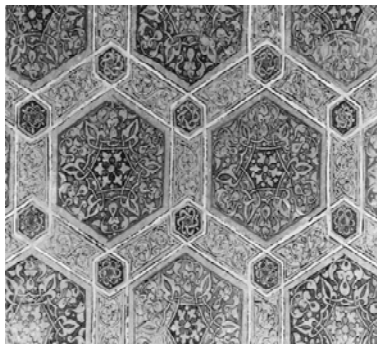
Coding Process

- 2 independent researchers
- nVivo software
- Began with coding two verbal reports, then met and discussed
  - Made adjustments to initial coding, then coded 4 additional verbal reports
- Coding unit
  - 1 item from 1 verbal report is a unit

Results: Research Question 1

- The new ECD items are eliciting higher order historical reasoning skills beyond simply factual recall and identification.
- See examples.

Previously administered item



The type of wall decoration illustrated above would most commonly be found in a

(A) stupa
(B) mosque
(C) cathedral
(D) temple
(E) marketplace

Key: B

Example verbal report

The type of wall decoration illustrated above would most common- commonly be found in a…
OK, well I see geometric shapes and I remember that, um… uh, I think… uh, Islam, they um… forbid to… the images of saints and stuff in their mosques and so they lean toward geometric shapes, so I would go with something like that.
A stupa… I have no idea what that is.
Mosque, there you go.
Uh… cathedral… cathedral, churches… churches, they have the stained glass, so I think they would go with images of… saints and stuff like that.
Um… a temple… I don't know what kind of temple they're talking about.
Or a marketplace… I'm not sure why there would be a wall decoration in a marketplace, so I think I would go with mosque.

New ECD-based items

The next two items are based on the following passage.

- "Most peasant farmers and herders, who constitute the great majority of the world's actual food producers today, aren't necessarily better off than hunter-gatherers. Time-budget studies show that they may spend more rather than fewer hours per day at work than hunter-gatherers do. There exist many actual cases of hunter-gatherer societies who did see food production practiced by their neighbors, and who nevertheless refused to accept its supposed blessings and instead remained hunter-gatherers."
- Jared Diamond, *Guns, Germs, and Steel*, 1997

#1

Which of the following would best support Diamond's argument in the passage above?

(A) Agriculture massively increased human population levels by increasing average caloric intake and life expectancy.

(B) The main result of the adoption of agriculture was a trend toward environmental degradation that continues to this day.

(C) Premodern societies living in areas that had both abundant game and/or fish resources and suitable conditions for farming tended to practice hunting and gathering.

(D) Hunter-gatherers needed to continuously limit their fertility and group size in order
to maintain the mobility necessary for meeting their nutritional needs.

Key: C

Example verbal report

- OK, so… the point he's trying to make. Jared Diamond- Guns, Germs and Steel, so I guess… I guess it's about arms, basically. Most peasant farmers and herders, who constitute the great majority of the world's actual food producers today aren't necessarily better off than hunter-gatherers. That's interesting. Time-budget studies show that they may spend more rather than fewer hours per day at work than most hunter-gatherers do. OK, I'd agree with that, but their reward is much greater, it's more stable. Um, there exist many actual cases of hunter-gatherer societies who did see food production practiced by their neighbors and who nevertheless refused to accept its supposed blessings and instead remained hunter-gatherers. I guess that is true, but over time they either adapted or they just died out, because um, agriculture is more stable and it's more promising. So I guess, um, his point is still that, um, hunter-gatherer societies… they're not as bad as… you think they are, I guess, they're- it's still possible to survive in that kind of society…. Premodern societies living in areas that had both abundant game and/or fish resources are suitable conditions for farming tended to prac- to practice hunting and gathering.Oh, actually that does make sense. Um… because, in the passage I see that hunter-gatherers… um, they had neighbors who did, um practice agriculture, but they themselves did not want to, so agriculture was possible, but they still um, didn't want to change, so… I guess C would work because there was abundant game and fish resources suitable for farming… and I guess abundant game and fish would work with hunting and gathering, so I would go with C.

#2

Which of the following types of evidence would a historian find most useful in evaluating Diamond's argument?

(A) Evidence about improvements in the amount leisure time available to agricultural workers as a result of mechanization.

(B) Evidence about calorie yields per acre in early agricultural societies versus calorie yields per acre in hunter-gatherer societies.

(C) Evidence about the population size of early agricultural societies versus the population size of hunter-gatherer societies.

(D) Evidence about nutritional deficiencies among early agricultural societies versus evidence about nutritional and vitamin deficiencies among hunter-gatherer societies.

Key: D

Example verbal report

Evidence about improvements in the amount leisure time available to agricultural workers as a result of mechanization… would that evaluate Diamond's argument… improvements in the amount of leisure time, that may, because um… Diamond said that agricultural workers, they had to put in more hours, so if you see evidence that they actually have some free time, that would, um… that wouldn't support Diamond's argument, so that would help in evaluating it.Um, evidence about calorie yields per acre in early agriculture societies, versus calorie yields per acre in hunter-gatherer societies.Um… well that's saying, basically, which one is better: agricultural or hunter-gatherer. Now the problem is figuring out which one, B, C or D.
Let me just go back to A, evidence about improvements in the lei- amount of leisure time available to agricultural workers as a result of mechanization… um… result of mechanization… mechanization… I'm not- I'm not sure Diamond is really focused on leisure, I think he's trying to say, um… like, for the amount of work you put, do you get the same reward. So I think I'm going to cancel A. Evidence about calorie yields per acre… calorie yields… food production… calorie yields… that would make sense, I guess. [INAUDIBLE] they're not better off and the whole reason for that is to get food…
Um, evidence about the population size of early agricultural societies versus the population size of hunter-gatherer societies. That doesn't really… he- Diamond's saying who's better off, so I don't think population has much to do with that, so I'd cross out C. Um, evidence about nutritional deficiency among early agricultural societies versus deficiencies among hunter-gatherer societies. So I think it's going to be between B and…Which of the types of evidence would a historian find most useful in evaluating Diamond's argument… most useful… is Diamond right or wrong… [INAUDIBLE]
OK, so I would say B, I think, because it's just the amount of food that you're getting for the work you put in.

Results: Research Questions 2 and 3

Item	Intended Skill	Evidence of alignment with intended Skills	Evidence of Alignment with other skills besides what is intended	Difficulty drivers identified
1	Legacy	n/a	Contextualization (*N* = 2) Historical Argumentation (*N* =1)	Degree of familiarity
5	Historical Argumentation	6	Interpretation (*N* = 6) Comparison (*N* = 3) Use of Evidence (*N* = 3) Synthesis (*N* = 5)	Presence of a stimulus, Difficulty of the stimulus, Quality of Distracters
6	Historical Argumentation	5	Use of Evidence (*N* = 5) Interpretation (*N* = 5) Synthesis (*N* = 5)	Presence of a stimulus, Difficulty of the stimulus, Quality of Distracters

Results: Research Questions 2 and 3

- Large degree of alignment between intended skill and observed skill
- More than 1 skill was often present for an item
- 3 specific task models could be revised to better elicit and align with their skill (i.e., Contextualization, Synthesis, Comparison)
- Difficulty drivers identified were noted for each item
  - Degree of familiarity was the most common (Indicated for 9 items)
  - Quality of distracters was second most common

CollegeBoard
inspiring minds

Discussion and future research

- Encouraging results for redesigned AP World History exam
- Recommendations for improvement:
  - Revisit 3 task models
  - Consider use of difficulty drivers in item development
- Limitations
  - Resource intensive and time consuming
- Future research
  - Analyze remaining verbal reports

pkaliski@collegeboard.org