COGNITIVE ASSESSMENT OF SUBJECT MATTER: UNDERSTANDING THE MARRIAGE OF PSYCHOLOGICAL THEORY AND EDUCATIONAL POLICY IN ACHIEVEMENT TESTING

CSE Technical Report 317

Eva L. Baker, Marie Freeman, & Serena Clayton

UCLA Center for Research on Evaluation,
Standards, and Student Testing

March 1990

Introduction[1]

For at least a quarter of a century, educators and critics have raised conceptual and technical questions about standardized achievement tests (Strenio, 1981). And for the most part, the public and its policy makers have ignored these ululations and continued to believe in the accuracy and usefulness of these measures, dismissing technical concerns as abstrusely academic and teacher complaints, at minimum, as self-serving. However, recent reform efforts, stemming from *A Nation At Risk* (National Commission for Educational Excellence, 1983) and other dark reports of American educational quality, have directed renewed attention and investment in achievement outcomes. With the statement of national educational goals by the President in 1990 and the governors of the fifty states in 1989, and the President's promise to measure achievement in grades 4, 8, and 12, standardized achievement tests are about to become national educational policy. The consequences of error in test design and interpretation are inestimably higher than in the past, for such measures will exert dramatic control on the public school curriculum, on what tests are published, and on what is taught. Information from achievement measures must answer three questions: What is the quality of our students' achievement? How can achievement be improved? Why can't present tests do the job? For the purposes of accountability and instructional improvement, the vast majority of existing standardized achievement tests are wholly inadequate. They create the wrong expectations and incite inaccurate inferences in terms of policy action. They are inappropriate in at least three central ways: their underlying theory, their content, and their procedures. These assertions deserve at least brief elaboration.

The measurement assumptions of standardized tests rely on models based in theories of stable individual differences. These models posit a general construct such as mathematics ability or reading comprehension and require at least two conditions for its measurement: (a) substantial variation among people on the target test in order to differentiate scores, that is, scores on the 78th or 64th percentile are intended to reflect different levels of performance; and (b) stability of measurement for individual performance for accurate prediction. When mapped against the requirements for assessing an individual's educational improvement or the impact of systemic educational reform intended to assure all students' success, these instruments do not measure up. Reports from most standardized tests obscure the meaning of the test scores from the teacher, the student, and the public. We may know the relative position of individuals and school districts compared to other individuals or school districts, but we do not know what level of performance any given score describes. Further, even under the best conditions, educational reform has weak effects. So to detect change, progress toward national goals, for example, achievement measures must be created that are sensitive to minor, but real differences in performance. Tests should tell us who has changed in ability to perform particular tasks at described levels of expertise. Standardized achievement tests do not tell us what we should want to know.

The problem of interpreting these tests is amplified by the way their content is selected. A major problem is content sampling within a particular subject matter, such as history or mathematics. Most subject matter measures are

1 We wish to thank Tom Kerins, John Craig, and Carment Chapman, of the Illinois State Board of Education; Bob Hill, of the Springfield Public Schools; Lynn Winters of the Palos Verdes School District; and the many principals and English and history teachers who participated in or helped with this project. In addition, we would like to thank colleagues at UCLA who helped with various aspects of the study: Pam Aschbacher, Jamal Abedi, Joan Herman, Edys Quellmalz, Merl Wittrock, Simon Chang, Yujing Ni, Regie Stites, Kyung-Sung Kim, and Rebecca Frazier.

commercially available and are intended to be sold to school districts and states. To be competitive, testing companies must attempt to include a sufficient number of topics with broad appeal in any subject matter. A common result, as predicted two decades ago (Popham & Husek, 1969), is a content-curriculum mismatch, where the overlap on test content and curriculum varies by district, school, or classroom. This phenomenon has been documented in many specific subject fields, for example, in mathematics by Floden and his colleagues (1980). In practical terms, a mismatch means that certain topics that are untreated in the curriculum of given classrooms and schools will be included on the test. On the other hand, even topics emphasized in teaching may only be superficially measured because of time constraints. Both types of errors result in misrepresenting students' actual achievement. One solution to this problem has been to encourage teachers to adapt their instruction to match test content (a process called "alignment"), a course of action that cedes enormous and inappropriate power to the developers of such tests.

A second, more global content issue is created by the pressure to test in a relatively limited number of subject matters. Such choices have been made as a matter of course to save money and time as well as to constrain the number of measures on which public accountability will be based. As a rule, districts and states commonly select an essential core of subject matter, often the areas of reading and computational skills in mathematics. Teachers and school policy makers adapt instructional time to focus on the goals to be measured. One consequence of this adaptation may be a reduction in time for untested subject fields: foreign language, the humanities, the arts, and the sciences. This reduction occurs logically to focus resources on the accountable aspects of the curriculum, but also because of the widespread, pernicious belief that students must learn the "basics" before they can profit from exposure to other subject matters and more complex intellectual processes. Particularly for poor performing students, opportunities in a wide range of subject matter are foregone (Oakes, 1986). The result is obvious—an educational system with clearly constricted curricula, differentially skewed to limit the access of the already disadvantaged.

The constraints on administration and scoring of standardized tests also influence their impact on school learning. Tests have been developed with strict time boundaries, partly in an effort to reduce testing time, partly from historical, psychometric reasons as a result of their original purposes to differentiate among individuals. To obtain differentiated and reliable responses, it is better to limit test time and to expose students to many short items. More test items also mean more topics can be covered. Multiple-choice items are the most frequently preferred achievement testing format because they are time sensitive, permit responses to a relatively large set of items, have acceptable psychometric properties, and allow economic scoring approaches. How do these choices affect students? Multiple-choice formats exert undeniable control on school practice. The format frames how information is presented, learned, and retained. These tests assess learning in an artificial, decontextualized manner that is remote from how students learn or will apply knowledge in the future. These tests are likely to reduce student motivation to perform and are likely to inhibit transfer. Such formats also convey a false sense of objectivity and quantification of performance, and objectivity and quantification are high-priority attributes in our society.

If it is true that such tests measure content only partially represented in school instruction and use formats convenient for administration and scoring rather than for learning, the simple problem of showing "improvement" in achievement is difficult and daunting. If it is difficult to design programs whose effects are detected by accountability measures, educators have few acceptable options. They may persist in doing the best they can, but may continue to see public confidence erode when test scores do not respond to their efforts. They may react in ethically

questionable or unacceptable ways, for instance, selecting tests that seem easiest rather than those that most accurately measure valid educational goals (Cannell, 1987; Linn, Graue, & Sanders, 1990), encouraging inappropriate practice of items on the test (Shepard, 1990; Popham, 1990), or even falsifying test results. Schools may respond by offering training programs to develop test-taking skills and may confuse, once again, ends and means. How do students, the nominal object of our concern, respond? From the broad evidence, it appears with less interest and focus at best, and with active subversion at worst. Thus, with the best intentions, policy makers, compliantly supported by the public, require standardized achievement measures as the principal indicator of educational effectiveness and continue to deform the system in serious ways. Even a partial acceptance of this analysis raises serious questions about the quality of inferences we are drawing from standardized tests.

New Choice Points

The expectation that accountability measures will directly and productively influence student achievement is wildly optimistic. Their imposition influences broad instructional choices: how much time is allocated to various subject matters, and what particular topics are covered. But to affect important student performances, measures must influence a far deeper and dynamic level of instructional decision making. They must provide guidance and be sensitive to differences not only in what topics should be included in the curriculum, but also to the dirty details of teaching and learning, the instructional processes that differentially affect performance. Unless measures are ultimately sensitive to significant instructional choices, their impact on school improvement will continue to be marginal, periodically stunning policy makers who use terms like "stall" to explain the dysfunction between their own accountability fantasies and the actual utility provided by test results for day-to-day instructional planning. One answer has been to search for alternatives to the existing tests that will provide help to improve instruction. Unfortunately, this strategy has resulted in the propagation of tests functions with little linkage between them—for example, between diagnostic and accountability tests. We need measures that can provide information at the right level of detail to guide instruction but that will not divert large proportions of instructional time from learning tasks.

To meet the legitimate concerns for accountability and resulting instructional improvement, we require new approaches. It is time to break away from the inertia of present achievement testing practices, from the never-never land thinking that we can make schools better only by trying harder. We need outcome measures that simultaneously avoid the major deficits of standardized tests and provide trustworthy and useful achievement information. Critical attributes possessed by new cognitive approaches to testing are: (a) their focus on important and teachable learning processes, (b) the confidence we can place in their measurements, and (c) the appropriateness of cues they provide for instruction.

Cognitively Sensitive Assessment

If we start with the notion that tests should measure significant learning in a way that supports desired performance, we are immediately led to a reversal of present practice. Instead of having tests constrain instruction, assessment procedures should be constructed to map directly on significant features of learning. Through close observation, skilled experts can tell whether learners are making progress on a wide range of intellectual tasks. Our problem is to transmute the critical aspects of that observational process into procedures suitable for use in large-scale assessment. We must shift our view from the measurement of broad constructs to the assessment of important and described classes of cognitive learning tasks—knowledge acquisition, deep understanding, and problem solving. These processes must be assessed as they are embedded in various tasks and content domains;

however, our assessment strategies may attempt to capture attributes of performance that transfer across subject matter domains. In our CRESST project on assessing deep understanding of subject matter, we conducted research designed to transfer knowledge that developed in learning research and apply it to the problem of assessing the understanding of history. What will follow is a chronological description of the developmental history of our project, interpolated by discussions of the generalizable problems confronting developers of new approaches to assessment.

Project Goals and Plans: Developing New Criteria for Scoring Writing in History

Stimulated by articulate statements about the importance of knowledge of history by Hirsch, Kett, and Trefil (1987) and the dismal performance of American students on tests of historical knowledge (Ravitch & Finn, 1987), we decided to focus our attention on the measurement of history knowledge, specifically aimed at assessing a deeper understanding of history. We conceived of the problem for students as a comprehension task dependent upon their ability to generate or construct meaning (Wittrock, 1974) from provided stimuli and by activating students' prior knowledge. This approach contrasts with the conception of history knowledge as a single construct dependent upon the accumulation of separate pieces of knowledge. Consequently, we broadened our approach from the usual multiple-choice format, building on our research group's considerable experience in developing measures of writing skill (Baker, 1987; Quellmalz, Capell, & Chou, 1982).

Our initial idea was to attempt to expand the content quality scoring rubrics used to assess writing and to apply them to subject matter topics in the field of history. Extant content quality scoring rubrics have treated content in one of two ways: as elaborated detail that contributes to a good essay in holistic scoring; or as important, unique material dependent upon the particular topic presented the learner. This second conception guides approaches used in scoring Advanced Placement Tests in History (Vaughan, 1983) and in primary trait scoring in the National Assessment of Educational Progress (1990). In this topic-dependent approach, individuals with expertise in the assigned topical area meet and develop post hoc standards for the particular set of papers written. The benefit of this procedure is the development of scoring scales that are particularly appropriate for the topic assigned. However, that strength is at once a severe limitation: First, the level of specificity required to adapt scoring criteria to a particular topic inhibits their more general use for other, similar topics. Thus, every topic possesses a unique set of criteria. Combining such particularized assessments across a range of topics or over a number of years involves a complex scaling process, based on equating results for different topics. Among a number of flaws, a major consequence of scaling is the ambiguity of score meaning. A second limitation relates to the inferences for instruction that can be derived from such measures. If every topic requires a unique set of criteria, what guidance can be provided to the teacher to inform teaching processes to improve student performance? Only if the tasks and scoring criteria are made public—released by the test producers—can teachers guide students to meet such standards, and then only if the same tasks are used. The trick is to find the appropriate level of generality to describe criteria so they are simultaneously appropriate for the particular assessment topics and conceived in terms that can guide future instructional practice and assessment.

Goals

The goals of our assessment research in the measurement of deep understanding of history were: (a) to develop valid formats for eliciting students' thoughtful explanations about history concepts; (b) to create and validate content quality scoring criteria for students' responses; and (c) to explore these

developments in the context of large-scale assessment settings. A longer term interest is to communicate the test design characteristics so that they will be helpful to the design of effective teaching strategies.

Strategies

Target. In light of our technical expertise in writing assessment, our project focused on essay writing in history. We believed that the strong tradition for this type of task in history instruction would increase the chances, if successful, of widespread acceptance of new assessment strategies. We also determined from reviews of plans for state assessment activities that writing in social studies was planned for many of the more forward-looking state assessment enterprises (for instance, California, Connecticut, Illinois, and Michigan). Finally, we believed that the present approaches used in the scoring of content-focused writing were inappropriate both conceptually and practically for the dual purposes of measuring deep understanding in large-scale settings and providing inferences useful for instruction.

Plan. In order to verify the need for essay scoring systems to assess content quality, we first had to determine if content specific scoring criteria for history already existed implicitly in the scoring behavior of history teachers. If so, we would identify these criteria, train others to use them, and validate their utility. If not, we would explore the literature to infer criteria that might be used. Even though our goal was to develop scoring approaches with reasonable generalizability across tasks to facilitate instructional improvement, we decided to limit our studies severely. We planned to focus on a grade level (11th grade) and on a single topic area in history, for we wished to be sure our findings were well grounded in a defined context. If we were encouraged by our results, we planned to test the generalizability of the approach: for other subject matter areas, for the age ranges of students for whom the approach was useful, and for sets of administration conditions. In sum, we anticipated the development of broadly useful assessment approaches as we conducted initial research in a restricted environment.

Our first problem was to identify specific content topics and strategies for data collection that would allow us to explore the issues of content quality scoring criteria. One requirement was to assure that students had some previous exposure to the concepts we planned to assess so that they could respond to our tasks. We hoped to assign passages in commonly used textbooks for this purpose. To that end, we reviewed textbooks, literature on the teaching of history, and available curriculum guides to determine the topics and most desirable sections of secondary school textbooks appropriate for our experiments in measurement. Our review of textbooks led to unoriginal but nonetheless depressing results. For every topic we pursued, we discovered that secondary school texts presented relatively superficial treatments, without sufficient concepts and depth of supporting knowledge to allow the development of deep understanding. These views have been supported in the literature by Beck, McKeown, and Gromoll (1989), Sewall (1987), and FitzGerald (1979). We also consulted at length with the staff of the UCLA Center for the Study of Teaching and Learning in History, a collaborative enterprise of the National Endowment for the Humanities that brings together experts in history and curriculum.

Goal Redefinition

Because we were unable to identify suitable text segments for use in the assessment, we decided to incorporate the reading of a provided text as part of the assessment procedure itself. This decision transformed in a serious way our assessment focus. Rather than an exclusive focus on measuring the accumulation of information developed over a long period of instruction, we now attended to two

major content issues: students' ability to read and integrate new information with previously learned knowledge, and students ability to explain new ideas using their prior knowledge. This transformation placed our work squarely in line with cognitive views of language comprehension (Anderson, Spiro, & Anderson, 1978; Rumelhart, 1980; Brown, Bransford, Ferrara, & Campione, 1983; Kieras, 1985). However, we were still driven principally by our subject matter concerns, a fact that guided the formulation of criteria for the topic and text selection for assessment tasks displayed in Table 1.

Table 1

Criteria for the Selection of History Texts to Assess

1. Must be a regular and significant piece of the secondary school history curriculum in the United States.
2. Must depend upon primary source material rather than summaries in textbook.
3. Must allow for multiple interpretations and inferences.
4. Must transcend immediate events and allow students to find relationships to other historical and contemporary events.
5. Must be brief enough to read within a class period.

Based on the application of these criteria, we decided that original speeches or essays composed by historical figures would meet criteria two, three, and five. For our initial set of studies, we selected the texts of the Lincoln and Douglas debates on popular sovereignty and slavery, choices that met the remaining criteria as well.

Identification of Content Quality Scoring Criteria: The First Pass

Our goal was to assemble valid criteria to assess understanding of history content. But essay writing consists of both content expertise and communication skills. We were well aware and troubled by the high intercorrelations in the literature between subscores on essays of expression skills and content knowledge (Baker & Quellmalz, 1980; Langer, 1984). Although it was obvious that highly verbal students would usually learn more about verbally based content areas, we were especially interested in discriminating performance between the ignorant facile writer with little subject matter understanding and the knowledgeable student with less developed writing skills. This desire corresponded to the common practice of high school teachers, who give both a "content" grade and a "form" grade (e.g., A-/B) on student essays. We wanted to focus on the elements that compose the content score.

A related concern was the impact of content knowledge (or lack thereof) on the raters' application of scores. We believed that knowledgeable people with experience in the subject matter would be needed to make the levels of distinction in which we were interested. Our first empirical study attempted to determine if the quality of content in essays, its accurateness, aptness, and structure, would be judged similarly by history teachers using implicit but common criteria for quality. We would contrast their ratings with those given by English teachers, specifically

We would contrast their ratings with those given by English teachers, specifically teachers trained to score essays in terms of the quality of general writing skill or expression, such as organization, style, and purpose. The essays we collected for this study were provided by 85 eleventh-grade Advance Placement (AP) history students in a suburban high school. We chose AP students because they would be likely to write "scorable" papers, that is, produce a sufficient quantity of writing to be graded. The AP students also had been exposed to an instructional sequence on the pre-Civil war period approximately five months earlier, so they would possess some background knowledge of the topic.

The experimental procedures spanned two consecutive days. On the first day students were given a general multiple-choice examination in pre-Civil War history, a test that had been validated by six expert history teachers. Next, students completed a background questionnaire describing their grades in English and social studies, self-estimates in ability, interest in writing and in social studies, and descriptions of teachers' instructional and assessment practices in history. On the second day, students were randomly assigned to read either the Lincoln or the Douglas debate text. After the students completed their reading, they were given an essay question in either a brief or an extended form that asked them to explain the author's main issues and why they were important. Students were allowed 50 minutes to read the text of the speech and to write their essay. The papers were independently scored by two groups of raters: the English teachers and the history teachers.

Procedures for English teacher raters. One rater group was composed of four English instructors, all highly experienced in rating student essays according to holistic and analytic techniques. All had been trained to use the writing scoring scales developed at UCLA (Smith, 1978; Quellmalz, Smith, Winters, & Baker, 1980) and subsequently adapted for use in numerous state assessments, research studies, and the international comparisons of written composition performance (Baker, 1987). These scales included four major categories—general competence, essay organization, paragraph coherence, and support (meaning detail)—as well as scales for grammar and mechanics. We also were interested in the thought processes that raters used and their initial levels of stringency. Thus, we asked raters prior to their training to read three sample papers privately, to rate them on a five-point scale, and to comment on their decisions and impressions; comments were tape-recorded. Raters also were asked to identify criteria for a good paper. The training was conducted using procedures described by Quellmalz (1986) with model papers and illustrations of score points. The raters were told explicitly to focus on issues of presentation and rhetorical effectiveness rather than content-specific issues, such as content accuracy and depth of explanation. Nonetheless, during the training the raters insisted on modifying the scoring system: They decided to include as part the general competence subscore some index of the student's attention to the specific writing task. All raters independently scored each of the 85 papers.

Procedures for the history teachers. Independently, and with no knowledge of the English teacher group or their resulting scores, a group of five history specialists was assembled to rate the same set of essays. Two were high school Advanced Placement teachers (from a school different than the data collection site) and three raters were advanced graduate students in history. Like the English teachers, all history raters were asked to assess three essays and to think aloud into the tape recorder as they completed this rating task. Their actual rating instructions differed dramatically from those given to the English teachers: No preexisting scoring scale was used, and no extensive training was conducted to determine if the history group shared implicit criteria. Each rater was told to give each paper two scores. The first score was to reflect how well the essay demonstrated serious understanding of the debate text read by the student. The second score was to provide an estimate of the essay's general quality, taking into

account issues other than the essay's content. These scores conformed to the content-form scoring mentioned above. We also asked the history group to select the ten best and ten worst essays, so that we could infer from their choices the operational criteria they used to make their judgments. Each history teacher independently rated each of the 85 papers, giving each a content quality and an overall quality score. Following the rating session, all teachers discussed in a group the attributes that distinguished the highest from the lowest rated papers.

Findings and Interpretations

Detailed data analyses were conducted; only the highlights will be reported here. No significant differences on student performance were found for text passage (Lincoln or Douglas) or question type (brief or extended), in the ratings of either group. Our findings verified the inappropriateness of the existing UCLA scoring scale for the content focused task we used. Alpha coefficients among raters ranged from a low of .52 for mechanics to a high of .75 for general competence (the one score where raters took into account the task content). This finding reinforced the need for the development of a content quality scoring rubric. For the history raters, the alpha coefficient on general quality was .69 and on content quality was .75. The generalizability ratings for English raters (4 raters by 4 subscales) was .65 and for history raters (5 raters by 2 subscores) was .73. An interesting finding was that the percentage of exact agreement for scores given in the history group to content quality was only 33%, suggesting that no clear set of implicit criteria was operating among the history specialists. In addition, a *t* test was computed between average scores given by the history teachers and the history graduate students; significantly higher scores were assigned by the secondary school history teachers. The correlation between general competence scores assigned by English teachers and history content quality scores on the same papers was .80, similar to the relationship between general competence assigned by the English group and the general quality score assigned by the history group (.82). Such data suggested that English and history teachers were looking at papers in fundamentally similar ways.

Unfortunately, the expert knowledge possessed by history teachers did not seem to differentiate their judgments of student essays. But some aspect of special knowledge was operating, however faint. A low but significant correlation was obtained between the content quality scores of the history teachers and the total multiple-choice knowledge score ($r = .32$, $p < .05$). Leads for the development of content quality scoring criteria had to come from other sources. We then reviewed the history raters' think-aloud ratings and their post-rating discussions of the ten best and worst papers. The historians agreed that the best papers had the qualities listed in Table 2.

Table 2

History Specialists' Generation of Criteria

Established historical context
Presented a sound thesis early in the paper
Detail contributed to thesis, was correct, and was not simply opinion
Avoided absolute judgments
Presented multiple points of view
Avoided interpreting the past in terms of present conditions

Scoring Criteria: Pass Two

In an effort to explore the utility of these criteria, a comprehensive and detailed scoring rubric was constructed based on these categories. The 12-category scoring scheme comprised the elements in Table 3 below; these elements were to be used as scoring dimensions for the papers.

Table 3

Scoring Criteria Inferred from Ratings of History Papers

Identification of the Historical Problem/Central Concept
Depth of Elaboration
Breadth of Elaboration
Flexibility
Fluency/Detail
Evidence of an Analytical Problem
Goal Orientation
Logical Structure
Evidence of Historical Analysis
Autocriticism
Presentation
Style

Detailed descriptions for each of five scale points for every category were prepared. Based on a brief tryout with raters and reviews by experts, however, we deemed that this comprehensive set of categories was too ambitious. A review of literature on characteristics of expert knowledge (see Voss, 1978) suggested how we could pare the set down to five categories thought to represent critical attributes of historical thinking: Historical Context, Depth of Elaboration, Breadth of Elaboration, Evidence, and Historical Analysis. In addition, we added two categories related to expression, Rhetorical Structure and Mechanics, as well as an overall quality rating, General Impression. New scale point descriptions were generated for each of the eight categories and model papers were assembled to illustrate particular attributes for training purposes. Four history raters (three AP history teachers and one history graduate student) were trained in the use of the new system. They spent two days rating the same set of 85 eleventh-grade papers used in the first study. Raters were observed as they scored papers and were queried about their satisfaction with the rating scales and training procedures. Raters had been given the scoring rubric in two forms: an extended, multipaged form with detailed explanations about each score point for use in training; and an outline of the dimensions. It was expected that after the initial training period the raters would use the outline form. However, they chose to continue to refer to the extended form, more rigidly adhering to the rubric than we expected. Raters reported that they could differentiate among categories and that they could also distinguish among criteria for score points (1-5) within each category. Raters were highly satisfied with the scoring categories and claimed to use similar criteria to score papers produced in their own classrooms.

Data from the second round of scoring were then analyzed. Unfortunately, the findings from these ratings did not significantly advance our research goal. Percentage of exact agreement among raters nudged up to about 35 percent, but alpha coefficients for rater agreement dropped to around .45. Most disappointing were relatively high intercorrelations (in the .80 range) among rating categories. These strong relationships were confirmed by a factor analysis that produced only two factors, one factor consisting solely of the mechanics rating and the other loading all other categories. These disappointing results forced us to regroup intellectually once again. Fortunately, we were able to compare the results from the first set of ratings by the five history teachers with this set of scores, since the identical student papers were read by both groups of history specialists. The categories in our revised rating scale that mostly highly correlated with the overall content quality rating from the first experiment were Historical Context, Breadth of Elaboration, and Depth of Elaboration; these categories were set aside for future exploration.

We so far had investigated the existence of common implicit criteria for content quality ratings, had analyzed the think-aloud protocols of raters, and had noted criteria used in identifying successful student papers. We then had created a comprehensive list of content-relevant elements, had reduced them to a smaller set of categories for feasibility purposes, and had trained a satisfied group of raters. Yet, we had not seemingly made much progress toward our goal. At this point we realized that our entire process had been guided in large measure by what history specialists said they valued and usually focused upon when they graded papers. It became obvious that such descriptions might reasonably be influenced by the raters' desires to appear to be comprehensive and thoughtful—in other words, by the social desirability of their answers.

Scoring Criteria: Pass Three

A new strategy for developing scoring criteria was employed, using the model derived from expert-novice comparisons (see Chi & Glaser, 1980, for an illustration). Rather than focus on what experts said they did, we were going to study their actual

performance on tasks identical to those provided the students. Three expert historians who were advanced graduate students in history, three secondary school history teachers, and three Advanced Placement students were asked to write answers to the same essay question used in the study above and to think aloud to permit us to assess their processes. The analyses of the essays produced by this process as well as our analyses of the think-aloud transcripts resulted in some clear direction for us in the area of criteria generation: Our analyses showed that all experts and some teachers used the elements in Table 4 to construct their essays.

Table 4

Elements Used by All Experts and Some Teachers in Essay Construction

A strong problem or premise that directed a focused answer
Use of prior knowledge, including principles as well as facts and events for elaboration
Text references (i.e., Lincoln speech)
Explicit effort to show interrelationships

In contrast, very bright but relatively inexperienced students and some teachers leaned heavily on the text in two ways. First, they often simply paraphrased or even restated the text in their answer. Second, they tried to cover all elements discussed in the text and were unable to distinguish between more and less important details. As a result of this analysis, a scoring scheme was developed that included all of the elements in Table 4, augmented by an overall general impression score. We were ready for new data collection

Rethinking Our Task

The first major redirection of this project occurred because of the paucity of quality textbooks and resulted in turning this assessment research toward the dual goals of measuring understanding and knowledge acquisition in the context of a particular subject matter corpus. The expert-novice analyses reshaped our focus in a second major way. If we accepted that prior knowledge in subject matter was essential to both premise-driven and elaboration components of quality of understanding, then it was clear that we should design our assessment situations to include explicit supports to enable students to access such information. We believed that we could do this in any number of ways and decided to explore a range of options, details of which we will expose below. More importantly, we perceived that this decision dramatically revised our view of assessment. We decided that the assessment situation itself should help students to perform the best that they could. We had moved into the blurry territory between learning and testing.

Revising the situation. Our next step was to create new questions to relate to the class of expert behaviors we had proposed as criteria. We decided to have all students read both the Lincoln and the Douglas texts to permit them to use comparison as a rhetorical structure. We developed two variations of essay questions, or prompts, which we experimentally crossed: One treatment condition included a narrative context for the prompt and asked the student to imagine being

in the pre-Civil War period and focus on an imaginary cousin as the audience for the essay; the other prompt presented the task as a more typical school assignment with the teachers as the implicit audience. A second set of treatments varied the instructions given to the student to assist their access to relevant prior knowledge. Although both conditions explicitly directed students to use their previous understanding and knowledge about the historical period in answering the essay question, one condition asked a series of stepped, short-answer questions to be completed before the student began to write the essay (see Table 5).

Table 5

Sample Prompt: Narrative Version

Topic:

Imagine that it is 1858 and you are an educated citizen living in Illinois. Because you are interested in politics and always keep yourself well informed, you make a special trip to hear Abraham Lincoln and Stephan Douglas debating during their campaigns for the Senate seat representing Illinois.

1) Unlike other tests, we hope you really will try to imagine yourself in the historical period of the debates, so take a couple of minutes to describe yourself, your family, and your work. (Spend about 2-3 minutes.)

2) As a well-informed citizen , you are aware of the many important events, laws, and court decisions that relate to the debates. List as many of these as you can. (Spend about 3-4 minutes.)

3) List, if you can, some principles that underlie our form of government and that are relevant to the debate. (Spend 3-4 minutes.)

4) While listening to the debates, you begin to think about the major problems confronting the nation. Some of these problems relate to principles upon which our government was founded. List the major problems you can think of. (Spend about 3-4 minutes.)

5) After the debates, you return home to find your cousin from England who has come to the U.S. for a visit. Your cousin asks you about some of the problems that are facing the nation at this time. Write the answer that you would give to your cousin, telling him/her about at least two problems that you feel are important. You can write this either like a regular essay or like a story. Just be sure to give your cousin the clearest picture you can. You may use any of the information you've identified above in your answer.

Be sure to describe each of the problems clearly and tell your cousin about events, laws, court decisions, and major principles of U.S. government that are related to the problem. Also explain the different solutions that are proposed to the problem, and give an example of what might happen if these solutions were adopted.

As a conclusion to your paper, write a brief summary that integrates the two problems and states your own position on the whole topic.

We also developed a prior knowledge test, basing it on the broad model developed by Langer (1984), for two purposes. First, we wanted to help students access relevant prior knowledge; second, we wanted to look at the relationship between that measure and rated use of prior knowledge in the essay. This 20-item test was created using a set of specifications to control the nature of the content queried. Students were to write brief descriptions or definitions for each of the terms provided, some of which were facts and events (e.g., Dred Scott decision), and some of which were at the principle (or at least concept) level (e.g., sectionalism). A few terms were irrelevant to the passage, and some were only tangentially relevant.

The new test administration sequence required two days. On the first day the students were to complete a personal information form (including details about their interests, age, etc.) and the 20-item prior knowledge measure. They then were to read the Lincoln and Douglas text segments and complete a short (14-item) multiple-choice test on information in the speeches. On the second day, they were to receive the essay question, write about 45 minutes, and complete a short debriefing questionnaire that asked for their reactions to the testing and for their estimates of their performance on the set of tasks tested. Following a pilot test in two Los Angeles classrooms, we tried the new assessment package in twelve classrooms in Springfield, Illinois.[2]

The Illinois Study

The purpose of the Illinois study was to test the assessment procedures under large-scale assessment conditions and to obtain data to bear upon the validity of our findings. Here we have space for only a short description and discussion of this study. In brief, 250 students in 11th grade participated, equally assigned from AP, college preparation, and regular classes. Two full class periods were allowed for the assessment. Students were told they were participating in a UCLA study to develop new measures for history. Since there were four treatment variations (stepped essay prompts/short prompts/narrative context/school context), students received their packets assigned at random within each classroom. On-site observers from UCLA administered the materials and collected information from teachers about their views of students' relative strengths in history, reading, test taking, and writing, and information about each teachers' instructional efforts in the topic area. In addition, we collected data from transcripts that reported students' course experience, grade point averages, and standardized test scores in writing, social studies achievement, and reading comprehension.

To obtain results, prior knowledge scoring rubrics were developed and applied to student responses. Scores ranged from 3, a fully elaborated answer, to 1, an incorrect or incoherent response. Two graduate students were trained to use the prior knowledge rubric and achieved .96 interrater reliability across the total measure (individual item agreements for the 20-item measure ranged between .70 and .96; 15 items had at least .86 agreement and only 2 fell below .80). Essays were rated using the new scoring rubric presented in Table 6.

---

[2] A good place for prior knowledge on Lincoln and Douglas.

Table 6

Elements of Cognitively Sensitive Assessment Scoring Rubric

Problem Focus
Prior Knowledge: Principles and Facts
Text Reference
General Impression

This time our empirical results were encouraging. Interrater reliabilities for the essay subscales ranged between .85 and .98. Intercorrelations among subscales were found between .0 and .60, supporting the premise that different aspects of student content quality were being assessed. Our findings also shed some light on the validity of the rubric. First, we determined that the measures reflected the different ability levels of the sample, with AP students scoring twice as high as the slower students on prior knowledge measures and on overall essay scores, and more than three times higher on use of principles in the essay. Our findings also showed strong relationships between teacher judgment of overall student achievement in history and our data ($r = .42$ for essay, .63 for prior knowledge). Our measures and standardized tests correlated .73 and .43, a variation based upon standardized test content.

Scoring Criteria: Pass Four

We reviewed our findings and decided it was time to test whether regular history teachers could be trained to use the cognitive scoring scheme. We also decided to revise the scale in a number of ways: to add categories for misconceptions and interrelationships, since in our own discussions we had not found a place in our system to take such concerns into account; and to refine the scale points for principle and problem focus. The categories in the scoring rubric are displayed in Table 7.

Table 7

Cognitive Assessment Scoring Rubric (1989)

Presence of Problem Focus
Prior Knowledge: Principles
Prior Knowledge: Facts and Events
Text
Interrelationships
Misconceptions
General Impression

We then conducted a training session with four high school history teachers to test the feasibility of our modified scoring approach. The training took approximately four hours, followed by the scoring session. Once again, we were very encouraged by our results. The prior knowledge measures and the essays were found to be reliably scored by teachers. Slightly lower interrater agreement overall was found for the high school teachers compared to the level obtained by project research assistants (alpha = .93 instead of .96). The interrater reliabilities on the essay subscales for teachers were in the .80-.90 range, except for the newly added misconception category (.68). Correlations between the prior knowledge measure and related elements of the scoring scheme were all reasonably high, averaging around .59, except for misconceptions (-.20) and text material (-.28). We conducted a factor analysis on essay subscales, and two major factors emerged. One factor included overall scores on content quality, the use of principles-based prior knowledge, premise-focused writing, and interrelationships. The second factor included misconceptions, the use of facts, and the use of text-based material. Although we are not completely convinced that this factor structure is sensible, the configuration of elements as it relates to the cognitive construction of meaning (factor one) and of the application of disconnected, and perhaps incorrect, information (factor two) is provocative.

Next Steps

Research subsequent to the Illinois study has been undertaken to verify the utility of the scoring system across topics, age ranges, and test administration conditions. We are looking at the performance of 9th-, 10th-, and 11th-grade students in two school districts. Data have been collected and are presently under analysis using two additional assessment topics. Both of these topics are drawn from the pre-Revolutionary War period and include texts by Paine, Henry, and Inglis. In addition, new materials have been developed for an extended assignment that involves Long and Roosevelt texts from the Depression period and incorporates as well additional resource materials for students' optional use. We anticipate a total of five hours will be needed for the assessment.

Limitations and Cautions

We have recounted the details of this effort to provide some insight into how assessment systems might be developed to reflect better the ways students actually learn and integrate subject matter material into their repertoires. We detailed our troubles and dead ends to demonstrate that the process of developing new kinds of useful and valid achievement measures is difficult and time consuming. New approaches to assessment are essential, but their development must be grounded in a theoretical view of learning. Establishing the validity of such new measures is also a difficult proposition. At least three major problems exist. One difficulty is the circular nature of new test development. Measures need to relate to but not be too strongly predicted by existing measurement strategies. A second problem with "deep understanding" tasks is the clear lack of systematic experience for the average student. Most students reported to us that our tasks were unusual for them. Their overall performance levels were exceptionally poor. To determine if our measures are truly valid (that is, if they reflected the desired class of learning), experimental studies must be constructed where students are trained explicitly in the process of integrating specifically presented material with various types of prior knowledge. Third, and most difficult, an optimal level of generality for task descriptions and scoring criteria is needed. This level must be sufficiently detailed to control raters' scoring behavior and to be valid for specific tasks. It must be sufficiently general to provide cues for teachers to use in planning and implementing instruction. A rough approximation of how such information can be economically displayed is provided in the specifications presented in Table 8. Such specifications would be augmented by detailed scoring rubrics with scale point definitions and also by a set of student papers illustrating, on different topics, various levels of proficiency. Clearly, a new program of psychometric research is needed. In the interim, we suggest that validity studies include criterion analyses by experts, experimental training studies, multiple measures of student learning processes, and demonstrations of statistical and conceptual connections to other reasonable estimates of performance, even including standardized tests.

We know that tests have driven instruction in the past. Can tests of the sort we are developing do so in a productive rather than a destructive way? What evidence do we have that teachers of history focus on the integration of new knowledge with prior information—the view that learners construct meaning? Are such tasks within the capability of all students? When we are constantly bombarded with stories that students don't know where the Pacific Ocean is or the half-century in which World War I occurred, is it naive to think that they can accumulate knowledge and use it to make inferences and explanations. These questions must be pursued. We believe that there are specific next steps to be accomplished. A major challenge is the development of a new theory of test design and validation, one that emphasizes individual learning rather than individual differences. Test designers must recognize that the measurement of significant processes takes significant time, and consequently tests of many short items and broad content sampling may need to be supplanted or supplemented by fewer more complex assessment situations. We need to develop concepts that will allow teachers to understand how to use such measures as an integral part of their instruction. Finally, we must get ready for the serious task of educating policy makers and the public about new models of assessment. We must counsel patience and anticipate that results are going to look worse, especially with new challenging measurement approaches, before they look better. When improvements eventually occur on cognitive measures such as those we have explored, we want them to reflect real and trustworthy learning for all students.

Table 8

Specifications for Writing Tasks

Discourse Type

Informative writing

Subgenre

Explain/infer

Major Cognitive Process

To demonstrate the acquisition of new knowledge or concept by contextualizing and elaborating position using prior knowledge (principles and facts)

Writing Process Measured

Drafting

Audience

Imaginary, peer

Topic Range

Subject matter based

History: A summary of major position by opposing statesmen

Information Given in Prompt

History: Text of speeches or essays written by historical figures (e.g., Lincoln)

Format

Brief text

Prior knowledge cues: Consisting of appropriate and inappropriate terms for specific processes, facts, or principles

Table 8, continued

Amount:

2 or 3 pages (no more than 10 minutes of reading)

A list of 10 to 20 entries for prior knowledge

Criteria

Content:

Organizing premise

Explicit use of prior knowledge, principles and facts (either provided or student generated) to explain or elaborate

Avoidance of misconceptions

Structure:

Relevant text references

Show interrelationships using text and prior information

Administrative Conditions

Time: 45-60 minutes

Resources: Students may refer to text and prior knowledge list during essay preparation

Interaction: None

Sample Prompt

Segment of Patrick Henry's speech, plus list of prior knowledge measure

Read the speech taken from the period just before the American Revolution. You are supposed to explain to a cousin visiting from Canada what Patrick Henry meant and what led him to the position he is in. Use help from the list of information to provide a clear answer.

Parallel prompt

Same except pre-Civil War, Stephen Douglas

References

Anderson, R.C., Spiro, R.J., & Anderson, M.C. (1978). Schemata as scaffolding for the representation of information in connected discourse. *American Educational Research Journal, 15,* 433-440.

Baker, E.L. (1987, September). *Time to write.* Paper presented at the Annual Meeting of the International Education Association, New York.

Baker, E.L., & Quellmalz, E.S. (1980, April). *Issues in eliciting writing performance: Problems in alternative prompting strategies.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Boston.

Beck, I.L., McKeown, M.G., & Gromoll, E.W. (1989). Learning from social studies texts. *Cognition and Instruction, 6*(2), 99-158.

Brown, A.L., Bransford, J.D., Ferrara, R.A., & Campione, J.C. (1983). Learning, remembering, and understanding. In P.H. Mussen (Ed.), *Handbook of child psychology, vol. 3.: Cognitive development* (pp. 77-166). New York: Wiley.

Cannell, J.J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average* (2nd ed.). Daniels, WV: Friends for Education.

Chi, M.T.H., & Glaser, R. (1980). The measurement of expertise: Analysis of the development of knowledge and skill as a basis for assessing achievement. In E.L. Baker & E.S. Quellmalz (Eds.), *Educational testing and evaluation: Design, analysis, and policy.* Beverly Hills, CA: Sage.

FitzGerald, F. (1979). *America revised: What history textbooks have taught our children about their country, and how and why those textbooks have changed in different decades.* New York: Vintage.

Floden, R.E., Porter, A.C., Schmidt, W.H., & Freeman, D.J. (1980). Don't they all easure the same thing? Consequences of standardized test selection. In E.L. Baker & E.S. Quellmalz (Eds.), *Educational testing and evaluation: Design, analysis, and policy.* Beverly Hills, CA: Sage.

Hirsch, E.D., Kett, J., & Trefil, J. (1987). *Cultural literacy: What every American needs to know.* Boston: Houghton Mifflin.

Kieras, D.E. (1985). Thematic processes in the comprehension of technical prose. In B.K. Britton & J.B. Black (Eds.), *Understanding expository text: A theoretical and practical handbook for analyzing explanatory text* (pp. 89-107). Hillsdale, NJ: Lawrence Erlbaum Associates.

Langer, J.A. (1984). Pre-reading plan (PRep): Facilitating text comprehension. In J. Chapman (Ed.), *The reader and the text.* London: Heinemann.

Linn, R.L., Graue, M.E., & Sanders, N. (1990). *Comparing state and district test results to national norms: Interpretations of scoring "above the national average"* (CSE Tech. Rep. No. 308). Los Angeles: UCLA Center for the Study of Evaluation.

National Assessment of Educational Progress. (1990). *The nation's writing report card.* Princeton, NJ: Educational Testing Service.

National Commission for Educational Excellence. (1983). *A nation at risk.* Washington, DC: U.S. Government Printing Office.

Oakes, J. (1986). Keeping track, part 2: Curriculum inequality and school reform. *Phi Delta Kappan,* October, 148-154.

Popham, J. (1990, January). *Appropriateness of teachers' test-preparation practices.* Paper presented at a Forum for Dialogues Between Educational Policy Makers and Educational Researchers, University of California, Los Angeles.

Popham, W.J., & Husek, T.R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement, 6,* 1-9.

Quellmalz, E.S., Capell, F., & Chou, C. (1982). Defining writing domains: Effects of discourse and response mode. *Journal of Educational Measurement, 19,* 241-258.

Quellmalz, E.S. (1986). Writing skills assessment. In R.A. Berk (Ed.), *Performance assessment: Methods and applications.* Baltimore: The Johns Hopkins University Press.

Quellmalz, E.S., Smith, L.S., Winters, L.S., & Baker, E.L. (1980). *Characteristics of student writing competence: An investigation of alternative scoring systems* (Report to the National Institute of Education). Los Angeles: UCLA Center for the Study of Evaluation.

Ravitch, D., & Finn, C.E. (1987). *What do our 17-year-olds know?: A report on the first national assessment of history and literature.* New York: Harper & Row.

Rumelhart, D.E. (1980). Schemata: The building of blocks of cognition. In R.J. Spiro, B.C. Bruce, & W.F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 33-58). Hillsdale, NJ: Lawrence Erlbaum Associates.

Sewall, G.T. (1987). *American history textbooks: An assessment of quality.* New York: Educational Excellence Network.

Shepard, L. (1990). *"Inflated test score gains": Is it old norms or teaching the test?* (CSE Tech. Rep. No. 307). Los Angeles: UCLA Center for the Study of Evaluation.

Smith, L.S. (1978, November). *Investigation of writing assessment strategies* (Report to the National Institute of Education). Los Angeles: UCLA Center for the Study of Evaluation.

Strenio, A.J. (1981). *The testing trap: How it can make or break your career and your children's futures.* New York: Rawson, Wade.

Vaughan, A.T. (1983). *Grading the advanced placement examinations in American history.* Princeton, NJ: College Entrance Examination Board.

Voss, J. (1978). Strategic difficulties in summarizing texts. *Reading Research Quarterly, 19,* 404-426.

Wittrock, M.C. (1974). Learning as a generative process. *Educational Psychologist, 11,* 87-95.