REPORT ON CONTENT DEFINITION PROCESS IN SOCIAL STUDIES TESTING

CSE Technical Report 310

Ernest R. House
Nancy Lawrence

University of Colorado

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

January 1990

The research reported herein was conducted with partial support from the U.S. Department of Education, Office of Educational Research and Improvement, pursuant to Grant No. G0086-003. However, the opinions expressed do not necessarily reflect the position or policy of this agency and no official endorsement by this agency should be inferred.

This paper was written as part of a research project sponsored by the UCLA Center for Research on Evaluation, Standards, and Student Testing (CRESST). It was prepared for presentation at the Annual Meeting of the American Educational Research Association, San Francisco, March, 1989.

Please address inquiries to: CSE Dissemination Office, UCLA Graduate School of Education, 405 Hilgard Avenue, Los Angeles, California, 90024-1521

Background of the Study

The Problem

Social science has been identified as one of the areas of highest priority for future testing. Greatly increased testing is expected at the secondary level in the next few years. One of the primary problems identified is the lack of an adequate analysis of the core skills and concepts within these content areas (Aschbacher, 1986). Linn, Graue, and Sanders (in press) have also demonstrated that the test performance of groups and individuals differs significantly depending upon the specific content of a test. The exact content contained within a test is of the utmost importance, yet the process of definition of test content has not been thoroughly explored.

As part of a larger study of content assessment, we are investigating what key concepts and ideas in the social sciences should be tested at the secondary level. In our original definition of the problem, we assumed that there should be congruity between the critical concepts in the social sciences on the one hand and the objectives and content of secondary school tests on the other. However, on reflecting about the matter, additional questions arose: Is there actually substantial agreement among social scientists regarding the key concepts in their disciplines? Should these key concepts, if identified, be taught to secondary students? If taught, should they be tested? What content is now being tested?

To put the matter another way, we had assumed originally that the secondary school curriculum was derived or should be derived from the social science disciplines. Certainly, that is how we think of the natural sciences and mathematics. Content in the natural sciences and mathematics at the secondary level should reflect the content in those disciplines, most experts believe, and one would be appalled if bad science or bad math were taught or tested. Of course, there are disagreements in all these disciplines about exactly <u>which</u> concepts should be taught at the lower levels, as well as <u>how</u> they should be taught. The relationship between the disciplines and the secondary curriculum is not as obvious in the social studies, however. In social studies in particular one might imagine that ideas quite different than the key concepts of the social sciences might legitimately be taught at the secondary level. The secondary school is the place where civic virtues, democratic ideology, and ideals of citizenship are emphasized in the curriculum. These notions are not necessarily identical to the main ideas in the social sciences.

Hence, the question of test content leads to broader questions about the social studies curriculum itself. Such reasoning required a broadening of our original study to include what is currently being taught and tested in the social studies areas. We examined the relationship between what <u>is</u> and what <u>should be</u> taught and the degree to which the tests should reflect this content. In the first phase we broadened our investigation to include not only the conceptions of what the social scientists see as critical and necessary but also what social studies teachers and social studies educators see as critical.

The Overall Design

The purpose of this content assessment project is to investigate what social studies content should be tested on national standardized tests and how that content should be defined. The content of these tests often seems to be taken for granted, or at least considered as essentially no problem, by many who see the important testing problems as shaping specific test items according to proper test statistics. However, it is clear that a test could have good items according to

professional standards of test development and yet not contain the content one might want.

Recent concerns about the content of what American students know and should know have been raised by critics such as E. D. Hirsch in *Cultural Literacy* (1987) and Ravitch and Finn (1987) in *What Do Our 17-Year-Olds Know?* Hirsch claims that there is a common core of knowledge that all Americans should know in order to be "culturally literate," while Ravitch and Finn contend that American high school students are abysmally ignorant of basic knowledge in history and literature. These and similar criticisms raise questions about what we should be teaching and testing.

The first phase of the project was to determine what the key concepts are in social studies. Is there a consensus among professionals as to what should be taught and tested? In the first phase we interviewed 16 historians, political scientists, social studies educators, and high school social studies teachers as to what they thought the key concepts are in their respective subject areas and what they thought should be taught in the public high schools in social studies. The research questions posed included these: What are the "key concepts" of the disciplines as conceived by these various groups? Are there important differences as to what disciplinarians, social studies educators, and teachers consider important concepts? The point of this first phase was to estimate the extent to which professional consensus could serve as a rationale for content definition. The findings appeared in our 1987 report (House et al., 1987).

The second phase was to relate these findings to the concept of "cultural literacy" (i.e., should we be teaching a common cultural core of ideas, and, if so, what should those ideas comprise and who should define them). This second phase required becoming familiar with the literature on cultural literacy. In the first phase we identified key concepts in the social sciences and in the second phase we examined the cultural literacy rationale for defining social studies content. These findings are in our 1988 report (House & Emmer, 1988).

The third phase of the project addresses the question of what the content definition process actually is in constructing major tests of the social sciences. We investigated exactly how content is defined in social studies testing by interviewing the developers of prominent high school social studies tests. The research questions included these: Is there a common process for defining content for assessment? How rigorous is it? What kind of content do these procedures produce? What is likely to be excluded? The findings from this phase of the project are the subject of this report.

In the fourth and final phase we will combine all our findings on definition of content for standardized tests with the results of similar work at the Center for Research in Evaluation, Standards, and Student Testing (CRESST), as well as the work of the other R & D centers, and make recommendations about what the proper processes for defining the content of tests should be and possibly what some of the content itself should be. These recommendations should be generalizable to other subject areas to some degree. The purpose of this report is to describe and analyze the actual process of defining content as currently practiced by major test developers.

Methodology of this Study

The third phase study reported here addresses the empirical question of what the content definition process actually is in the construction of standardized social studies tests. We investigated how content is defined in the social studies testing area by interviewing the developers of several prominent tests, including

those involved in several National Assessment of Educational Progress (NAEP) history and social studies tests, a major commercial test, and a prominent state assessment.

Our method of investigation was to contact the test development administrators, interview them regarding the content definition process, then ask them to identify the people most closely concerned with the content definition tasks in social studies, both inside and outside the testing organizations. We interviewed those people and asked them for names of other participants specifically involved in the content definition effort. For example, since the overall procedure employed for the NAEP history tests was to have a committee develop objectives and write items, and then to turn the consequent item writing tasks over to the NAEP staff, we interviewed people inside and outside NAEP. Altogether we interviewed 10 people who were involved in developing several different social studies tests.

From these interviews and from written materials about the development of the tests, we constructed a realistic portrait of exactly how the content was defined. Our research questions included: What was the actual content definition process? What kind of content does this process produce? What was excluded? Our data collection procedure was to develop a structured interview protocol to elicit answers to the research questions and to tape-record the responses. The interview protocol included general and specific questions about the content definition process. In sum, we asked participants for a reconstruction of the content definition process.

We also examined a major state test and a major commercial test. Like the NAEP tests, these are premiere examples of the tester's art at the national level. Eventually, we contrasted the content definition processes and the type of content that each process produced. We promised anonymity to participants interviewed so their responses could be more honest. A description of those interviews is listed in Appendix A, and the interview protocol can be found in Appendix B.

Findings of this Study: The Process of Defining Content

The NAEP Tests: The 1988 History Assessment

The initial content for the NAEP history assessment was defined by the learning area committee, a group of historians and social studies teachers who represented different regional areas, genders, and racial groups. The group contained one geographer and one political scientist, but no sociologists, anthropologists, or economists. It is not clear how members were chosen for the committee. Some members had worked on the 1986 history assessment. The committee members were sent preliminary mail packets and information about various state guidelines for social studies. The chair of the committee, a history professor who had worked on the U.S. History Advanced Placement Exam, was assigned the task of determining what content domains should be covered. He did not want to be interviewed.

At the first meeting in Princeton, the committee was instructed to define specifically what content should be tested at given ages. Starting with the 1986 NAEP history exam for reference, the committee decided on content in 2 two-and-a-half-day meetings:

> There were relatively few questions on the [1986] exam from social or economic history. The questions on economic history tended to be largely about inventors. It had a few questions about women's history and a number of questions about black history on the exam. Not

much about immigration—one or two questions about that. We went through that exam, looked at the questions and looked at the balance in the exam. So we began with that as one of the sources of questions and one of things to look at in conceiving what topics ought to be included. We then had a lot of other questions. ETS staff had assembled quite a lot of questions from a number of sources at ETS that were produced for the American history achievement test, for instance. So we looked at a lot of questions, and then we spent quite a lot of time talking about how to divide up the territory of American history so that all the major topics would be covered and would be covered in appropriate ways for the audiences to whom the tests were being administered. (Interview D)

The 1988 committee divided American history into chronological periods and into topics within those periods, including themes in political, constitutional, economic, social history, and the history of particular groups, such as women and racial and ethnic groups. A draft of content was prepared by the ETS staff and sent to committee members for review. Members embellished the draft:

> Okay, so this committee was called together, and the first meeting is, after preliminary mail reviews and exchanging of information of different state guidelines, any source of resource materials that are available through mail review, the committee was to draw together. I forget the exact date, but at the first meeting, it was to kind of come up with some sort of, at least consensus from the committee view on what should be tested or what should be learned, what were the goals for American students at the given age. From that meeting draft texts were prepared by ETS staff. The draft test was reviewed, further embellished by members of the committee, who would also write sections as well as review it. (Interview C)

After all was said and done, the domains agreed upon for the 1988 test were those listed in Appendix B. There was no problem achieving consensus in the learning area group, and this ease in attaining consensus on what was to be tested is a general theme. However, in the 1988 panel, there was some disagreement with what had appeared on the 1986 test:

> That [1986] was a smaller committee. We were a little more (and I think Ravitch is quite explicit about this) inclined to be prescriptive. There are things people should know, and there had been debate over whether they've identified the correct things and whether those are the things that people are teaching. When we looked at what they [1986] had done, we felt that they had omitted a lot of things that people do teach and that we had thought were important. One of the things we looked at in putting together the assessment exam that NAEP had just conducted was previous exams that NAEP had done in the field of social studies. I think there are two such exams. I don't know whether it's beyond the scope of your project, but you might want to look at those. What was fascinating was that those two exams were entirely different from each other and were entirely different from what we did. (Interview D)

The much earlier NAEP social studies test had contained little history, so there was no basis for comparison with either the 1986 or the 1988 assessment:

> Those tests had no history in them at all. We could not, one of the problems we had was there was no baseline for determining whether kids now know more or less history than they used to. There is a

> baseline math. There is a baseline now in reading level or in writing. There is no baseline in history. Because the previous test didn't do that. As I recall, one of the tests was primarily, well I thought, most of us on the committee thought, was deeply objectionable, although it had been administered nationally under NAEP auspices. The reason I thought it was objectionable was it was full of questions that were highly valuated. The test assumed there was a correct answer. The students were asked, "What would you do in this circumstance?" or "Have you voted," for instance, "for student council?" And the right answer was "yes." It was a series of tests about right behavior as a citizen in America. And right beliefs, and there were certain beliefs, many of them I think were widely shared, about fairness and non-discriminatory attitude, but those were correct attitudes, and that was the way the test was set up. So it was a test to find out if people were good citizens and the report to Congress would be "most Americans are good citizens, or not good citizens" or "moderately good citizens." The other was a test that was kind of all over the lot, with regard to a smattering knowledge about civics and American political institutions, a smattering of knowledge about your neighborhood and so on. It was hard to define what the content was. And it wasn't comparable to the test whether you were a good citizen or not. So NAEP has done, I believe, three assessments in the general area of social studies and they're entirely, each is entirely, incompatible with the other. (Interview D)

The NAEP staff understood its task as creating consensus on the assessment content. Some of the academics had difficulty with this push for consensus, but they eventually accepted the process:

> We were there as advisors. NAEP's job was to organize and orchestrate a series of discussions that would produce, in the end, a consensual product. Their job was to identify matters that might be controversial and to get, make sure the issues got aired and to isolate those items on which there wasn't any agreement, but to see whether in fact you could reach agreement by defining issues in ways that were mutually acceptable all around and so on. (Interview D)

The learning area committee helped develop the test items based on the content booklet. Some initial items were constructed by the learning area committee itself, additional items were written by outside writers hired by ETS and by ETS staff, and ETS supplied items that had been developed for other tests. In two meetings the committee reviewed hundreds of items for accuracy and wording. Items were matched to objectives:

> Items were constructed. We used the committee themselves. We used outside item writers. They would be edited for style. They would be reviewed. We have staff here [ETS] who are specialists in social studies. The staff would prepare booklets with these items for committee review. We would have a meeting. The committee would have seen the items beforehand. They would have rated beforehand. And at the meeting, the committee would go through item by item and review each item. If the mail review had shown that clearly everyone had not liked the item, they would just tear it out. If some people liked it and some people didn't like it, and if they thought it could be fixed, they would massage it at the meeting. And then if everybody liked it, they'd look at it and say, "How about using this one?" And then they would use it. Then after that meeting, we had a pool. (Interview C)

The items were assembled into blocks requiring 15 minutes of testing time and field tested. Again the committee met and reconsidered the items with the field test results. Items were eliminated if no one knew the answer or if they were insufficiently reliable:

> And then we field tested all the items. We sampled across the country. We did a national field test. So that we had kids from all different areas of the country responding to these items. Okay, after the field test, we had item statistics. We prepared booklets again of all the items that were field tested and with their statistics and sent them off for committee review by mail. We then brought the committee back here to ETS and we had a big meeting. And again, we had their ratings by mail. We then sat down at the meeting and those items that either had bad statistics or the committee finally on a second look at those we didn't like, we threw them out. Those items that the committee liked and had good statistics were kept for the assessment. At least they went into the pool. After that meeting was finished, we had a pool of pre-tested items that had good statistics from which test developers here at ETS [text missing in transcript] social studies tested be assembled again into 15-minute blocks and these blocks were paired, not paired, but put into groups of three to four booklets, and they were arranged so that each block appeared without the other blocks in different booklets. So we had a balanced and completely blocked design. We made many booklets with these. They were spiraled in the classrooms across the country and that constituted the national assessment, and that's how the items were prepared and administered.
>
> [Interviewer:] What criteria were employed for items the test?
>
> Suitability. Okay, as I have mentioned earlier, matched to the objectives. Kind of a rethinking of the objective itself. Okay. Dealt with items that we wanted the objectives to get to. I mean it's testing, what we hope it's testing. And is that, is this important for kids to know? And did the items show good discrimination on, did it have good statistics? When I say that, were students who did well on the entire tests also doing well on those items? Or were kids who were doing very poorly on the test, doing well on this item? What we wanted to do, what we try and do, is get items that good kids are getting right and items that bad are getting wrong. (Interview C)

The actual writing of the items would be done by an item writer. The interviewer asked if test items were ever rejected:

> Yes. When those items are sent to them, they would say, "This one stinks." And it's confusing or it's too general or it doesn't do what you want it to do. Or it isn't appropriate for the grade level. Or it has some bias in types of elements with it, in terms of women or blacks or whatever. So there would be those criteria to say that I think you need to get back to the drawing board or modify them.
>
> [Interviewer:] Now, could one person wave the red flag?
>
> Yes. (Interview E)

A few items were eliminated for political reasons:

That [bias] was high, prominent in the minds of everybody involved. I mean, Ina [Mullis, deputy director of NAEP] was very much aware of that and Walt [MacDonald, NAEP director of test development] and the committee I was on certainly [were] very concerned about that, and that was, I'm sure, a concern of the state coordinators as well. And the process was designed to include representatives of minority groups, and women, of course, were prominently involved in it. So, there were those kinds of questions. There were, there may have been other kinds of questions about the political implications of asking about certain things. Assessing Republican administrations or whatever. I don't recall that that was ever directly flagged. There were certain kinds of questions that some, there was some discussion, that maybe this isn't the kind of thing that Office of Education is going to want examined, but as I recalled that, they may have raised such a question only about one or two items altogether. There were hundreds of questions. And I didn't, quite a lot of this was informal discussion. (Interview D)

One of the main constraints was the very limited amount of time that the committee had to work:

You know, that kind of thing happened, and so we tended to be, to give fuller consideration to whatever it was we took up first and by Sunday morning we'd be racing and that happens every time you have a committee meeting of that kind. In general, it was, nonetheless, it was, I don't mean to say that things were terribly rushed or things got badly neglected. The one area, in the end, we didn't get many questions into the assessment on items of American culture and literature, and I think that was partly a difficulty in finding questions that you could get a reasonable amount of kids who would know the answer. And partly a matter of just that was the last of our priorities and we didn't get to it. Our emphasis turned out to be on political history, a certain amount of treatment, considerable treatment really, to social history, less of economic, quite a lot of attention on the political history of constitutional matters and then, that is partly a result of the fact that there was a parallel civics knowledge, and that had a lot of questions on the role of the court, the structure of American national government—and much of that is constitutional—and the role of the appeals process. There was a good deal on [the] constitution [and] legal matters in the civics examination or assessment as well as in the history one. And so and we worked on both of those. So I think we, you know, my memory of it is there was a lot of politics and constitution but a lot of that was really quite appropriately in the civics part. (Interview D)

The learning area committee never saw the final test, although some were involved in putting together the final report of results. In the view of one participant, the NAEP process reflects current social concerns, which change from time to time:

The NAEP process is designed to reflect the current concerns of the people involved all the way from, on the one end from, or various ways to characterize it, but including scholars and people involved with managing states and regional school systems and people in the classroom. And I think there were other groups the materials were shown to. I think, perhaps, some parent groups and others who were also, exactly how many and how and what degree of input they had,

> you'd have to get that from NAEP. Now, there's no question that an agenda was set in Washington in this case, that worked close, of course, with Ravitch and Finn, and others. E. D. Hirsch was involved in that initial *What Do 17-Year-Olds Know?* [Ravitch & Finn, 1987]. That group was concerned with classic texts in American political history and the core of national history as they define it. The test that emerged is not the test they would have written, but it was a history test. And you know, historians thought that that was okay. (Interview D)

After the primary committee defined the content objectives, a second group representing 40 states reviewed the objectives. At least 50 to 60 people received the document for review. The Department of Education received the document, and they in turn sent it out for more review. This review by hundreds of people was incorporated into the history objectives document. Content consisted of eight major historical periods (e.g., exploration and colonialization):

> We received things ahead of time, probably we were given, I would say, a month or so to respond with our suggestions. And then we met together for two days at NAEP headquarters. Then the learning area committee, the ones that were setting up the parameters, looked at our suggestions and comments. Staff worked on modifications and sent them back to us in the mail for the advisory committee members to comment on, and then we sent it back to the staff and then they proceeded from there. I would say that the process for our receiving correspondence and state advisory committee meeting, receiving revisions sometimes and sending it back, probably [lasted] over a period of several months at least (Interview E).

There was considerable agreement on the content among the state level committee, but less agreement on the type of test items covered:

> There were differences in opinion, I think, in terms of the degree to which what was being put together would be assessing rote types of knowledge rather than students' ability to comprehend at higher thinking levels and to apply or demonstrate that they understood what was being presented about, shall I say, the rights and responsibility of government. So, there would be more disagreement about the test of assessed items or the process being a lead into a traditional type curriculum and instruction. (Interview E)

The NAEP Tests: The 1986 History Assessment

The preceding 1986 history and literature assessment was a special case of NAEP in that Chester Finn and Diane Ravitch had a grant from the National Endowment for the Humanities to assemble and conduct an assessment that fit Finn's, Ravitch's, and William Bennett's idea of what content should be, what some might call the "cultural literacy" approach. Nonetheless, the assessment followed the outline of most NAEP assessments in terms of its process. The learning area committee was selected by Finn and Ravitch, and some had served on previous NAEP assessments. The history committee consisted of one ETS staff member and five others, including a state agency person, two local school district persons, and two historians, one from Princeton and one from Harvard. Ravitch and Finn were also on the history committee, but spent most of their time on the literature committee.

At a three-day meeting in Princeton, committee members were instructed to define the content that everybody should know—people, events, important dates,

important documents. They started at the beginning of American history and listed the chronological periods. Disagreements were usually decided by including the disputed material. In all this deliberation the ETS staff acted as facilitators, maintaining only a background presence, as was the case with the 1988 history assessment. The biggest problem was with the modern period because it was not certain what material the students would have covered. Textbooks did not play any overt role in defining this content. The sole criterion of selection was "items of information that should be in the heads of Americans" (Interview A). So the 1986 assessment provided more specific and somewhat different instructions to the primary committee than did the 1988 assessment.

Recalling and listing facts took most of the committee's time, and unlike the 1988 history assessment, the committee did not write items. For example, on the NAEP geography assessment, the geography committee wrote half the items, and ETS produced the other half from their files; items then were reviewed by the committee. However, on the 1986 history assessment, ETS staff apparently wrote all the actual items. As in 1988 there was a state review of the material. A committee of people from state agencies reviewed the content domains already specified:

> Okay, basically people that I would classify as experts in the area of U.S. history had identified the content, and there was a list of objectives classified into what I call domains or broader content areas and so forth. We looked at that and basically reacted to it, filled in the gap. The attempt initially was to identify almost all the content that people thought was important for students to know basically across all grade levels. And my role in that was basically to react to it, and I offered some of ideas of things that I thought should have been emphasized and so forth. (Interview B)

Again, at this level there was no problem with consensus. The idea was to list everything the committee thought important, knowing that only a small part of that content could be tested. On the 1988 geography assessment the learning area group reviewed the items one by one, with committee members raising questions or suggesting improvements in individual items on the spot:

> There was a meeting where we chose the items first, what we're going to field test. Then we were all given an opportunity to edit them through the mail. They were sent to us. I sent it back to them. Then they took the changes from how many people there were, I think there were six or seven of us, and they sent us another copy, and then they said, "This is the one that is going to be field tested. Is there anything that bothers you?" Then we would send that back, and then they would go ahead with field testing, and then we met again and we looked at the results after field testing and made a selection of the final items for the test. Then, they actually did it on the national sample and then we met again and analyzed it. (Interview B)

When all was said and done on the 1986 history assessment, the broad domains in which test items were written were those listed in Appendix B.

The Commercial Test

The commercial test that we examined was from a highly reputable publisher that had just developed a new version of its major social studies test. The process of defining content for the test differed considerably from NAEP:

> The process we use is something that goes on for about a year prior to the test item development, and, of course, we make a complete survey of state and major city, district, and diocese curriculum guides. We also looked at the major basals for the elementary schools and the major texts that we have determined are being used in the schools from a survey of publishers for the high school texts. Meantime, we're also trying to keep up in philosophy by reading the various journals and publications and attending meetings of NCSS and other meetings that are applicable. We have, prior to establishing an objective structure of any sort or any kind of approach to the test, we also have brought in some consultants from a local or state school and got an opinion from them as to the major trends. We have also a curriculum advisory committee which is made up of, I think it's eight or nine people from all over the United States that we are able to ask specific questions of. When they come, they spend a couple of days here every year. (Interview F)

The curriculum committee was composed of general curriculum people from school districts and university people, not content specialists. Their function was to identify trends. The test itself is written by staff hired temporarily for about six months or over the summer, ordinarily former teachers who have been item writers before. For this particular test, the staff took six months to identify content. The staff examined the documents collected in the survey and developed a taxonomy of content from that. History, political science, sociology, anthropology, economics, and geography were combined into a social studies test. What did this cover?

> Geography. Locational applications that were mainly map reading and interpretation. Regions and physical features. Human and environmental interaction. And in economics we tried to cover the fundamental concepts of economics. Economic communities. Resources and technology. Regional energy dependency and global interdependency. In history we had time cognition and chronology. Significant events and people.
>
> Patterns of civilizations. In political science we tried to cover governmental structure, democratic process, rights, responsibilities. Regional and global issues. Sociology and anthropology, individual and groups roles. Cultures. Belief systems. Multicultural perspectives. Issues and attitudes. And then under the objective called interrelated disciplines, that's where we dumped things according to their disciplines. If it was something that seemed to strongly link and emphasize the link between geography and history, we'd put it in that category. But these were items which emphasized the relationship rather than being based on the relations, rather than being based on factors from different disciplines, [these items] actually emphasized the relationship. Then the last one we applied to social studies. We more or less grouped those into assessing information, concepts, and using the information or using the context. (Interview F)

Again the members of the group had no problem coming to a consensus as to what should be included. The original content objectives were outlined by the content editor in consultation with the supervisor and the project director, who set up the original content structure. The staff then wrote items on these identified themes—about 8,000 accepted items altogether for different forms of the test. Item writers were given a short course in item writing and the objectives outline, and the items produced were eventually reviewed by at least 10 people, including 2 outside reviewers and a "bias review" by minority reviewers. More than half the items were

rejected following the field trials on about 500,000 students. The test took about three and one-half years to develop.

The State Test

The state test that we chose to examine is from a large state that was trying to develop a different approach. The initial history content for the test was defined by a written framework document, which was prepared by a committee with the help of outside consultants:

> The [1987] framework sets out various literacies that we expect to complete in our development of content, kindergarten through 12th grade. The literacies address three different goals, and one goal is the goal of knowledge and cultural understanding, and under that there's historical literacy, ethical literacy, geographical literacy, economic literacy and social/political literacy. The second goal is democratic understanding of civic values and the strands of that have to do with national identity and constitutional heritage, civic rights values and responsibilities. And the third goal is skills obtained in social participation. Strands that are under that are basic study skills, critical thinking skills, participation skills. The framework focuses on historical literacy and geographic literacy as the foundation stone for the framework on all grade levels. (Interview H)

The framework was drawn up by a committee of 19 people, with a large number of other people playing important roles as consultants and reviewers. The state agency selected the people to participate, and candidates had to submit a written statement of why they wanted to serve on the committee. Coming to agreement was not easy:

> The real question, I think, came as to how we teach these, and there was a camp of people that wanted to teach the social studies: geography, anthropology, archeology, political science, economics, history, etc. And the social studies, the organizer and thematic approach, was the persuasion of part of the members of the team. Another group on the team really felt that history needed to be reinserted as the curriculum driving force. History/social science—that was a real political struggle within the committee. Quite honestly there was one that cost us some time because the thematic approach verses a chronological/historical approach with history and geography the core. I think it was simply, just simply a power balance at the end that did it. That there were finally enough people persuaded to another direction, that there was a swing group of people, and they went in the direction of history/social science and history and geography, and just like anything else, I think that it had to deal with personal contacts and persuasion of the arguments. There was a group in the committee that was still dissatisfied, that still wanted to keep going back to a thematic approach that was loosely organized around social science and that faction just lost. There's no other way of saying it. (Interview H)

A 35-person committee then took the framework document and began defining content within the domains specified, dividing the framework into elementary and secondary committees. The committee started with the idea that perhaps the multiple-choice format was not what they wanted. Rather than develop a highly detailed outline of content, as they had done in the past, the elementary committee tried to determine what the most important ideas were, to

articulate how these ideas might be taught in a classroom, then to think how one might capture the classroom events in an assessment.

Although the basic structure of content was already defined, the elementary committee had a difficult time deciding what the important ideas were. The chair of the committee suggested that they identify stimulus materials, such as original documents, from designated time periods and develop test items based on those:

> After some months of trying that, the committee just threw up its hands and said, "We can do this no more. This is not something that we can do. We're not capable of this task." At this point there was a lot of frustration, just a lot of things that we didn't have answers to, so the next attack that we tried was to kind of revert back to the old style of kicking every content as articulated framework. And, for instance, if under one of the units it says, Beginning Civilization, the Near East, Mesopotamia, Egypt, Cush, we could develop a very detailed matrix. And if there was a mention of something, you know, we'd try to find or create an item or develop an item that reflected that matrix. That also proved very frustrating, so we were at a point where fine detail didn't seem to get at the content. Starting with [text missing in transcript] material didn't seem to get at the content. And we were floundering. What we ended up doing, and that was almost a question of expediency, was we found some interesting instructional materials and most of those instructional materials came from a summer institute. (Interview G)

The state agency brought one hundred teachers into a summer workshop, asked graduate assistants to "discover" historical documents, and developed test items based on those documents. The teachers first discussed the content and documents among themselves and then developed model lessons. The assessment committee based the test items on the lessons, the original documents, and the K-12 framework. The content for the test then came from the primary sources. An example of units:

> Okay, I have one in front of me that's from an archaeological text and what it reflects is, it's a series of dwellings of how a particular housesite looked at four different times. Okay. And then they have some description of that. Here's another one that came from a music source on how to, you know, how to make a lyre from a variety of sources. Here's one that's on Egypt. It's a collection of documents, translations of documents that says, "Advice to a schoolboy from a father." You know. "I place you at school along with children of notable..." You know, it's an original source document. Okay. Here's one. Discussing the economics of Cush. And the teacher has a lesson related to that. Here's some very nice poetry, some Chinese poetry that's been translated. I'm looking right now at a very large binder. Let me estimate the size of the binder to be 350 or so pages, xeroxed mostly on two sides. Fairly rich materials, historically and then these are organized into a series of lessons and the kinds of lessons, for instance, the one I first opened to is man and his artistic expression. It takes a four-week unit, has 10 major topics in it. (Interview G)

The teachers in the summer institute who developed these materials did not work from a set of objectives but from their own sense of exciting and interesting ways of approaching the topics. An agreed structure emerged as the teachers worked on materials (e.g., "In every unit let's reflect both the place and time, and let's reflect the institution, events, and major figures"). The slant came from their own interactions, from arguing and writing, not from the state:

> Now they did struggle, and understandably they struggled with trying to fit in, in a 36-week calendar year, the entire content of, sort of a sweeping content. How do you really cover some early humankind, sort of fall of the Roman Empire with sixth graders in 36 weeks of literature and history? I mean study literature and science and so on. And that was especially difficult at, say, seventh grade where we're talking about some very diverse cultures and a wide sweep. And they were constantly dealing with attention and saying, "Well, the best we can do is to give lots of rich things from which teachers may themselves choose." So the goal of the assessment thing became showing that indeed schools were often in richness. So, we've taken it as a watch word and knowing that assessment is very directly translated into structure to make sure that there was kind of a full gamut of activities and a strong emphasis on our open-ended problems. Problems that are really looking more like instruction than assessment. (Interview G)

The tests for fourth, fifth, and sixth grades were based on these sorts of materials and were written by about nine members of the state committee, including three teachers from the summer institute, mostly selected by the chair of the committee. Members who were expert in particular areas created items, which they then shared with four or five others. These items were then passed on to in-house reviewers for editing:

> But this gave us a set of really good materials and gave us item writers a sense of you know, from this content area, what seems to be really rich and important. And then we could use that for the basis for items, and we started on that track and just literally, the test fell together. I mean, everything just went very, very quickly. But until that time, we were floundering, and so it took this instructional emphasis to kind of heighten our attention and give us a sense what the ideas were and what the key notions were and that's what we're going to try to reflect. We're trying to reflect all of the strands. So, we'll need to create items about the economic, Neanderthal civilizations and so on. (Interview G)

After editing, the items were reviewed first by the history group within the state agency and then by grade-level chairs. After that the items went to outside reviewers, who checked for ethnicity bias and historical accuracy. A group of curriculum coordinators porvided another check. The items were judged either up or down, and many were rejected. Field testing of the items will be the next step, and these results will be examined at next year's summer institute. "So we're aiming at a more authentic model. We're aiming at a more classroom-based, more portfolio, you know those kinds of models are there. So, in that context we know that we want to send some messages about implementation of the framework. We want to send some messages about how important history is" (Interview G).

The former process of test development was rather different:

> The last process would first involve the assessment committee. The committee then would develop a very detailed outline and matrix. So we're talking in terms of very detailed outline. How that outline would be determined would have been from the framework, but also from, you know, textbooks and things like that. It was commonly used textbooks. And then from that matrix we would then write very specific items directed towards .4 B C 6, if you understand what I mean.

> The frustration of our committee is that we don't have that nice comfortable outline. We don't have a sense of completion and those kinds of things, and we're talking about working from bigger ideas. That's one major difference. A second major difference would be [that] all the previous tests have only been multiple choice. From the beginning we wanted tests that included much more than multiple choice. And there are folks on the committee who very, very strongly feel that we should have no multiple choice. There are people on the committee who feel very strongly that we should have all multiple choice. So the compromise at this point is some of each. We're clearly aiming at lots of other assessment types other than multiple choice, and we're field testing portfolios, open-ended problems, multiple-choice questions and essays—essays that would be read jointly by the direct writing folks and also by history folks. (Interview G)

This innovative state test is still under development, and it is still too early to see how it will come out. The secondary test committee pursued a more traditional approach, working from a detailed outline.

Conclusions

Participants. A relatively small number of people are engaged in defining the actual content of the standardized social studies tests. That is, many of the same people work on the different tests at the state and national levels. Some of the people we interviewed have worked on at least two or three tests. All the people seem dedicated and highly capable. Surprisingly, there are only a few subject matter specialists among them. Only one of the people we interviewed has a Ph.D. in history, even though history forms the core of most of these tests. Most of the work is done through committees, and the typical work committee has one historian. Other social scientists are not usually included. On the other hand, the state test did involve quite a number of people, if one counts the teachers who developed the instructional units.

Variety in process. The NAEP content definition process differs significantly from that of the commercial test and the state test. The typical NAEP process is to assemble a small committee and have that committee define the content, with checks of various sorts from reviewers. However, the NAEP process differs enough from year to year that the three NAEP social studies assessments conducted since the beginning of NAEP are not comparable to one another. By contrast, the commercial test developer relies heavily on an analysis of the content in the most widely used social studies texts. The state test involved a large number of people engaged over a considerable period of time, although we believe the large scale of this enterprise is unusual. For the most part, these processes do not produce the same content on the tests, and, in fact, the content on the three NAEP social studies tests is quite different. One might forecast with some confidence that results on these tests would differ considerably if given to the same population simply because of content differences.

Purpose of assessments. The four assessments differed in terms of their stated purposes. One of the assessments was developed specifically to complement the design of a new state-adopted history/social science framework that calls for a return of history to the core of social studies. The other three assessments were designed independently of particular history/social studies curricula.

NAEP (1988): "... an education research project... to collect and report data over time on the performance of young Americans in various learning areas" (NAEP, 1988, inside cover).

NAEP (1986): "The purpose of the probe is to gather information about basic knowledge in U.S. History in a fair, accurate, and replicable process and to make it available in an accessible, intelligible form to prospective users—a population that includes national, state, and local policymakers for education, teachers in the humanities, professional educators at every level, parents, citizens, and taxpayers" (NAEP, 1986, p. 7).

The state test: "... an effort to strengthen education in the history/social science curriculum. It is centered in the chronological study of history; history placed in its geographic setting. History is integrated with the other humanities and the social sciences. History/social sciences are correlated with other subjects such as language arts, science, and the visual and performing arts."

The different meanings of literacy. The term literacy is a buzzword in current assessment efforts, and it assumes different meanings across different content and assessments (NAEP, 1986). For example, the 1985-86 NAEP assessment produced the *Foundations of Literacy* booklet and acknowledged that "traditionally, educators have tended to define 'literacy' as a set of reading and writing skills rather than a body of knowledge, shared references, and commonly understood facts that enable people to communicate with one another. In reality, being 'literate' includes not only having the skills to communicate, but having some knowledge about the variety of topics that form the basis of dialogue and information-sharing, oral or written." This particular NAEP assessment "ended up with a master list of content that would define history" (Interview A).

In the 1987-88 NAEP assessment, the term "literacy" was dropped. It did not appear in NAEP/ETS booklets or during interviews. However, when asked if content definition is based on the assumption that there exists a common body of historical knowledge that students should know, respondents answered:

> Yeah, I think that's about the only way you can do it. You have to ask the question, "What is it you expect the students to know by the time they reach such and such a grade?" Initially the idea was to list everything that you thought was important, knowing full well that the final test would only measure a sample [of this knowledge]. At least initially it was to identify everything that people thought was relevant and important for students to know. (Interview B)

> Yes, historical knowledge and knowledge about government and politics. Specifics. We assumed there exists a body of historical knowledge kids should know. (Interview E)

> That was a distinction I think you would find between the committee that developed the assessment test for NAEP and to some extent the committee that was chaired by Ravitch and Finn. We were a little more inclined to be prescriptive. There are things people should know and there has been debate over whether they've identified the correct things and whether those are the things that people are teaching. (Interview D)

On the commercial test "literacy" was not mentioned, but one interviewee responded that content definition was based on the assumption that a common body of historical knowledge exists for high school students. On the state test literacies were defined as historical, ethical, cultural, geographic, economic, and

socio-political. For example, historical literacy includes developing a keen sense of historical empathy, understanding the importance of history as society's "common memory," maintaining its identity and transmitting its traditions and ideals to each new generation, and understanding the importance of religion, philosophy, and other belief systems:

> The literacies address three different goals and one goal is the goal of knowledge and cultural understanding and under that there's historical literacy, ethical, geographical literacy, economic literacy, and socio-political literacy. The second goal is democratic understanding of civic values and the strands of that have to do with national identity and constitutional heritage, civic values, rights, and responsibilities. And the third goal is skills obtained in social participation. Strands that are under that are basic study skills, critical thinking skills, participation skills. It is with a clear look at the three goals and a focus towards trusting each of the literacies and strands under the various goals that the test is devised. We're taking a look at how could we teach historical literacy, for example, while we are taking a look at civilization in Mesopotamia. What is included in historical literacy?... The framework focuses on historical literacy and geographic literacy as the foundation stone for all grade levels. (Interview H)

All in all, literacy turns out to be an elastic term.

Bias checks. Perhaps sensitive to recent charges that the SAT and other standardized tests suffer from racial, ethnic, and gender bias, each respondent underscored the bias checks in place in their respective assessments.

1985-86 NAEP: "There would be an overt attempt to include a female person in this committee and a minority, ethnic minority. It is my understanding that ETS has a normal way, have some checks on this and exactly how they do this I don't know" (Interview A).

1987-88 NAEP: When asked how the 1987-88 NAEP guarded against bias, one of the interviewees, who is a member of a minority, stated, "They had people such as myself. They definitely would have included along the way someone that was going to comment on the items being developed in terms of race, gender, ethnic group and religion, and we were asked to comment specifically on that area." NAEP also had a very strong concern about regional biases: "We tried to develop a committee that had regional representation for the country. We tried to get people from the East Coast, West Coast, the middle of the country, the South. We tried to develop a committee that also had gender balance. It had minority representation" (Interview C).

The commercial publisher: "We have a publication on guidelines against bias that has been published by our parent company, and we send that to the reviewers and they use that as guidelines. Reviewers have been selected from various minority groups, outstanding educators, and I think although we can't have all of them review every piece, there are representatives mainly from Black, Hispanic, and Native American groups. We also have at least one person who's a national authority on gender bias" (Interview F).

The state test: "Folks charged with reviewing items/content for bias—standard practice in state for 10-15 years. Every item has to be a fair yardstick...not a rubber measure. Review results in terms of major ethnic groups in state. We tried to balance our committee regionally and ethnically" (Interview E). Another respondent said:

> Plus, the way the committee was set up, we had people from the northern part of [the state], southern part of [the state]. We had people from, gee, we had [an individual] from World Affairs Council, so we had a world affairs approach. We had the Constitutional Rights Foundation...We had teachers, principals, curriculum...I mean it's mind boggling how they put the committee together. There were ethnic groups that were on the committee. There was a black teacher there from the elementary school...There were Hispanics on the committee. There was ethnic group balance. Balance in terms of geography from the state—people from the north and people from the south—and there were people, professors, and consultants, and I think,...it better addressed the diversity that we really wanted."
>
> [Interviewer:] Were there ever instances where a minority, a person representing a minority group, asked that more questions be included, addressed?
>
> Well, they had to make this kind of a question in writing. And then a copy of that was distributed to every member of the committee and that was one of the homework assignments, just to read that response. So [we] responded in small groups and then came back and discussed in large groups. People had, very clearly had, access to us. (Interview H)
>
> Construction of test items involved being sure to include a variety of history-social science concepts, geography skills, thinking skills levels, levels of complexity, and concepts and types of test items that would also appeal to girls...and students of various ethnic and socio-economic backgrounds. (Interview I)

The usual way of dealing with the bias problem was to appoint someone of that race, ethnicity, gender, or regionality to the panel of developers or reviewers, or both. NAEP seemed particularly concerned about regional biases and was careful to involve participants from the four regions of the country. Other kinds of bias, such as social class bias, were not attended to in any substantial way by the testers.

Reliance on texts and primary and secondary sources. The processes differed considerably in regard to how much they relied on textbooks and primary and secondary sources. The NAEP tests did not rely on textbooks at all in their content definition process. However, a careful analysis of the 1986 NAEP history test shows that almost all the content on the test is covered in the most commonly used high school textbooks. It may be that the process of having a small group of people meet for a few days and define what should be on the test reproduces what is in the texts. Many of the participants involved had authored or had used texts extensively. Or it may be that the professional NAEP item writers produced and winnowed items according to what is in the texts, though we found no evidence of this.

The content appearing on the commercial assessment seemed driven by what was in high school history and social studies textbooks. In contrast, history texts were also reviewed by the developers of the state test, but mostly to see if they contained primary documents. Developers of the state assessment employed research associates to gather primary and secondary sources from libraries, books, and other resources. These materials were collected and were made accessible to the individuals designing the curriculum and subsequent assessment.

Reaching consensus. One of the most puzzling findings of this study is the ease with which the participants claimed to have achieved consensus. Yet our earlier study (House, et al., 1987) indicated that there was little consensus among historians, social studies educators, and social studies teachers on what should be taught and tested, though there was consensus among political scientists. Our 1987 sample was small and confined mostly to one state, but it is not likely that consensus would increase with a larger, broader, and more diverse group of respondents.

Perhaps the consensus emerges from the content definition process itself. That is, the NAEP staff works hard to achieve harmony and agreement among those who define subject content. At the same time, they do not interfere directly, but rather act as facilitators. What happens with NAEP is that the learning area committees are selected by NAEP through a process that is not clearly specified; these few people have often worked on standardized tests before and sometimes together. There were few full-fledged content specialists among the members of the committee. These few were historians. Usually the committee consists of an historian or two, teachers, and curriculum specialists. Having only a few historians may reduce the possibility of disagreement since the other members are not likely to argue with the historian about content, and the possibility of disagreement among historians is reduced by having only one or two.

The internal workings of the process itself are also important. These few people are brought to Princeton for a weekend in a cloistered environment. They see themselves as consultants to NAEP, rather than as representing other interests or constituencies, and know they have the enormous task of defining the content of the entire test within a few days. Anyone who has served on a jury knows how a jury is encouraged to unanimity in a short period of time by the circumstances of performing such a task under constrained conditions and sometimes by fatigue. Having to do such a task under such circumstances induces one to rely on that content on which one can find agreement.

The recent NAEP history tests emphasized testing on basic historical facts on which everyone could agree and not upon difficult and controversial content. Within the NAEP process itself there was also the principle of deciding content on the side of inclusion. Where there was disagreement, the disputed fact was usually included rather than excluded, with the proviso from the NAEP staff that they could look at the item later to see how well it performs in the field trials. Hence, if disagreements do arise, there is a handy way of dealing with them. And ultimately many items are eliminated by poor test statistics. There are also some indications that controversial material might be excluded sometimes for political reasons and for the sake of reaching consensus. It is also the case that the NAEP staff themselves are excellent in dealing with people, as several participants mentioned. The NAEP staff know how to smooth out the rough spots and handle difficult situations.

Hence, our analysis suggests that consensus emerges from the careful selection of participants, the constrained situation under which they labor, the nature of the task itself, the careful management and human relation skills of the NAEP staff, and the use of the field trial as an adjudicating mechanism. There is nothing sinister about achieving consensus. NAEP itself was built on compromise and careful consensus building by Ralph Tyler, Jack Merwin (the first NAEP director) and others and would not exist without such consensus building. The question for test construction is, "How is the content changed as a result of these processes?" In other words, given all the content that might appear on a social studies or history test, what actually does appear and what is eliminated?

It is instructive to see what happened when some of these constraints were relaxed, as with the state test. In the state test development a larger number of people, 20 to 40, defined the original framework for the social studies in the state

over a period of a year and a half. Considerable disagreement arose over fundamental issues, such as whether social science themes should be pursued as opposed to a chronology of historic events. The chronological history approach was eventually adopted, mostly by a political vote. In addition, large numbers of teachers and participants who were brought in over six weeks during the summer, developed a number of unusual test items. In other words, with the relaxation of the constraints, consensus was not so easily achieved and the content defined for the test was significantly different.

The commercial test achieved consensus by its hierarchical development and control. A few people in the company managed the development of the test using textbook content and curriculum guides as the explicit content source. Item writers were hired to write to the content framework, so that the entire process was centrally controlled. The only place for potential conflict was with the external advisory board, but they seemed to act more as advisors than as reviewers. All in all, we judge that the content of what is on the standardized social science and history tests varies significantly according to the process used for obtaining consensus and that this consensus is likely to eliminate certain content (e.g., controversial material).

History verses social studies. There is a fundamental disagreement over whether social studies tests should cover history or the social sciences, reflecting perhaps a schism between history and social sciences in the academic disciplines themselves. For example, the 1988 NAEP test was history only: "The NAEP exam was really explicitly not a social studies exam. And that grew out of a concern that knowledge in history was by itself something that ought to be assessed and that was distinct from knowledge in social studies. [The committee] had a number of historians, a geographer, and a political scientist. It did not have sociologists, anthropologists or economists" (Interview D). Ravitch and Finn spent considerable time lamenting the intrusion of the social sciences into the public school curriculum, and the 1986 NAEP history and literature tests were designed explicitly to emphasize history subject matter that had been neglected, in their opinion, and to exclude social science content.

On the state test both the participants and the state framework emphasized the return of history to the center of the social studies curriculum, though not without considerable struggle (Ravitch was a primary consultant), including: (a) knowledge and cultural understanding, incorporating learnings from history and the other humanities, geography, and the social sciences; (b) democratic understanding and civic values, incorporating understanding of our national identity, constitutional heritage, civic values, and citizenship responsibilities; and (c) skills attainment and social participation, including basic study skills, participation skills.

Only the commercial test included social science content in a broader test of social studies: "History test not separate—part of social studies test" (Interview F). This test included history, political science, sociology/anthropology, economics, and geography.

Importance of chronology. The history included on these tests was not just any history but the history of chronological events, mostly political events. To those who advocate the return of history to the center of the social studies curriculum, the learning of basic facts, including important places, dates, names, and events, learned in chronological order, is particularly important. There are other ways that history could be taught, of course. The current trend in historical scholarship is to study social history, personal history, the history of economic institutions, the history of women, the history of childhood, and so on. This creates a history which may be chronological but which focuses on different events and subject matter than the traditional history of political and military events. Many

people who have been active in defining what should be on standardized achievement tests oppose the inclusion of this newer history into the high school curriculum.

Content varies with the times. The fact that the three NAEP social studies tests varied dramatically in content, even to the extent that they are not at all comparable to one another is an illustration of how social studies content varies directly with the current of the times. The test constructed in the 1960s was laded with items about good citizenship with little factual history content. By 1986 the reverse was true—all facts, no citizenship. The earlier content was more consistent with the views of the high school teachers and social studies educators we interviewed as to what is most important, and with the social protest movements of the 1960s. By 1988 this content was seen as arbitrary and even objectionable by the panel defining the 1988 content. One might conclude that some way of establishing continuity in content would be highly desirable, but that such content is highly political and subject to social pressure and changing social beliefs.

Need for a more systematic approach. Although we will integrate the various strands of our project next year, along with other work being done at CRESST and at other OERI centers, and will draw some general conclusions and recommendations about how the content definition process might be improved, it is not premature to suggest the direction of improvement. A great deal of effort is invested in sampling and item selection in standardized achievement testing and comparatively little effort in content definition, which is equally important. In our judgement, the current content definition processes are not faulty, but they could be improved substantially. They seem to us not as systematic as they might be, and we are concerned about what content appears on social studies tests from year to year. There does seem to be something a little arbitrary about it.

Selection of people. Generally a only few people were substantially involved in defining the content of these tests. There is nothing wrong with the people who were involved, but one wonders if bias creeps in when there are so few people involved. Many more people are involved in reviewing items, but that is a different enterprise than defining content. It seems to us that there could be broader involvement from a more representative group; especially more content specialists could be involved. Typically only one or two historians were used to help define the history content, for example, and one would think that a more representative group of historians would be useful in defining content that is likely to shape the national curriculum in that subject area. The same would be true of other subject areas, of course.

Time and interaction. Participants were limited to only a few days of deliberation in the NAEP process, literally a weekend or two. One would think that defining the content of the test for the entire country would deserve more time. The content selected does make a lot of difference in many ways, and this content varied considerably. Such a serious enterprise with far-reaching ramifications deserves more time and effort. One has the impression that the committees hurry through the weekend exercise doing the best they can, but also reaching decisions they might not make had they more time to consider and deliberate. Having longer and more sessions could improve the quality of the content considerably.

Aids and materials. The state testing program examined here represents how more interesting and innovative things might be done with enough time and materials available for study and stimulation. Admittedly, when compared to the more usual processes, the process they undertook was an expensive one. Theirs might not be the best model to pursue, but what they have done is certainly innovative and apparently successful in producing different items and different content than more typical procedures produce.

Content checks before, during, and after. After the content domains are established and items are written, field tested, and selected or rejected, there is no regular procedure for reassessing what content is actually going to appear on the test. What if an entire domain is eliminated because of poor items? There is no way of checking this and balancing the test against the original domains. In fact, the way these processes now operate, there are not enough checks for content accuracy and coverage. Checking individual items to see that they are good is not the same as checking to see what content is covered in a broader sense. In general, we think there should be more consideration of material by subject matter specialists, as there already is for regionality, ethnicity, gender, and race. To reiterate, defining the content on these standardized tests is just as important as any other phase of test development, and it deserves equal care, consideration, and effort.

References

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

House, E.R., & Emmer, C. (1988). *Cultural literacy and testing* (Report to OERI, Grant No. G-86-0003). Los Angeles: UCLA Center for the Study of Evaluation.

House, E.R., Emmer, C., Kolitch, E., Waitz, B., & Baker, E.L. (1987). *Definition of content in social studies testing: Conceptual content assessment report* (CSE Tech. Rep. 273). Los Angeles: UCLA Center for the Study of Evaluation.

Linn, R.L., Graue, E., & Sanders, N. (in press). Comparing state and district test results to national norms: Interpretations of scoring above the national average. *Educational Measurement: Issues and Practice.* [Also issued as CSE Tech. Rep. 308. Los Angeles: UCLA Center for the Study of Evaluation.]

National Assessment of Educational Progress. (1986). *Foundations of literacy: A description of the assessment of a basic knowledge of U.S. history and literature* (Description Booklet No. 17-HL-11, CN6710). Princeton, NJ: Educational Testing Service.

National Assessment of Educational Progress. (1988). *U.S. history objectives 1988 assessment* (CN6710). Princeton, NJ: Educational Testing Service.

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

Appendix A

Backgrounds of the Participants Interviewed

1. **Interview A:** Experience—NAEP Advisory Committee for the 1985-86 Foundations of Literacy Assessment (U.S. History); 1987-88 NAEP Learning Area Committee, Assessment of Geography

2. **Interview B:** Experience—1987-88 NAEP State Advisory Committee, United States History Assessment; 1987-88 NAEP State Advisory and Learning Area Committees, Assessment of Civics: United States Government and Politics; 1987-88 NAEP Learning Area Committee, Assessment of Geography;
   B.A. University of Wisconsin
   M.A. Geography, University of Minnesota
   M.A. Political Science, Penn State University
   Ph.D. Political Science, University of Wisconsin, Milwaukee
   - High school social studies teacher (8 years)
   - Geography and political science professor (5 years)
   - District-level evaluator in charge of test development in state agency for last 10-12 years

3. **Interview C:** Experience—NAEP
   B.A. Biology (Rutgers)
   Ph.D. Biology (Rutgers)
   - 10 years teaching sciences at university
   - Published in scientific and popular journals
   - Authored high school science texts

4. **Interview D:** Experience—1987-88 NAEP Learning Area Committee, United States History; 1987-88 NAEP Advisory Committee, Assessment of Civics: United States Government and Politics;
   B.A. Government (Harvard)
   M.A. Teaching (Reed College)
   M.A. History (Columbia)
   Ph.D. History (Columbia)
   - Public high school teaching experience (American History)
   - College/university teaching (History)
   - Published in *American History Review* and *Journal of American History*

5. **Interview E:** Experience—1987-88 NAEP State Advisory Committee, United States History Assessment; 1987-88 NAEP State Advisory Committee, Assessment of Civics; 1985-86 Foundations of Literacy, History Reviewer
   B.S. Social Sciences Education with emphasis in History (Ohio State)
   M.A. Social Sciences with emphasis in Sociology (Michigan State)
   Ph.D. Curriculum Instruction (Michigan State University)
   - Public school teaching experience at the elementary and secondary
   - Administrative experience in the public schools
   - Adjunct professor at state university
   - Currently, Social Studies Specialist with State Department of Education

6. **Interview F:** Experience—Commercial publisher project supervisor
   B.A. Social Sciences (Stanford)
   B.S. Social Sciences Education with emphasis in History (Ohio State)
   - Author of two young adult books
   - On and off with commercial publisher for last 20 years
   - Full-time with publisher last eight years

7. **Interview G:** Experience—State Assessment
   B.A. Philosophy (UCLA)
   M.A. Early Childhood Education (UCLA)
   Ph.D. Educational Psychology (UCLA)
   - University faculty member (Elementary Education)
   - State Department of Education for twenty years
   - Worked for commercial publisher for 2 1/2 years
   - Six and 1/2 years in public schools (elementary)

8. **Interview H:** Experience—State Assessment Program Advisory Committee
   B.A. Political Science/History (Cal State - Hayward)
   M.A. Educational Psychology (Cal State - Hayward)
   - Licensed psychometrist
   - Social studies public school teacher - 22 years
   - Writer and contributor to state K-8 Guide
   - Served on the state assessment program steering committee
   - Involved in 1988 History Social Science Framework (1 of 5 teachers of 250 that applied)
   - Recipient of the 1988 Teacher of the Year award
   - 2nd place winner of Curriculum Development award

9. **Interview I:** Experience—State Assessment Program Advisory Committee
   B.A. Math
   - Multiple Subjects teaching credential
   - Special Education (Gifted) teaching credential
   - Teacher of gifted students (K-6)
   - Teacher of highly gifted, lower socioeconomic students (5th and 6th)
   - State Geographic Alliance participant

Appendix B

Interview Schedule

Content Definition Questions

1. Would you describe in general the process for defining the content for the history/social studies test?

2. What specific areas of content were defined? Examples? Did people agree?

3. How did you go about identifying the content for the history test? Who was involved? How long did it take?

4. Was content definition based on the assumption that there is a common body of historical knowledge that students should know?

5. Were history texts used as the basis of content? Which ones? Were texts used for content item ideas?

6. Did you work from a set of objectives when defining content? If so, what were the objectives and who outlined them?

7. What criteria were observed when defining and selecting content? Were any guidelines followed? Examples?

8. How did you construct items for the test? Did everyone have to agree on the items?

9. To what extent did consensus exist regarding history content? How much consensus was needed before content was adopted?

10. How were test items proposed?

11. Why were test items rejected? Who made the "rejection decision"?

12. What were the standards by which a test item was judged? Who set the standards?

13. Were test items later pulled? Examples? Why were they pulled?

14. Were test items and content "tested" before appearing on the assessment? If so, were they tested on students?

15. Was there an "in-house" review of test items by individuals not directly involved with the content definition process? If so, who were the individuals?

16. Was content reviewed by individuals independent of the testing service? If so, who were these individuals and how were they selected?

17. Were test items reviewed for content after appearing on an assessment test? If so, how were they reviewed and/or redefined and by whom?

18. How long has the current definition process been in place? How does the current process differ from the last process? Why the changes?

19. How did you guard against potential bias to particular groups?

Appendix C

Content Domains of the Tests

A. NAEP (1985-86):
   I. Exploration and Colonization: up to 1763
      a. Exploration
      b. Colonization
   II. The Revolutionary War and the New Republic: 1763-1815
      a. The Revolutionary War
      b. Establishing the New Nation
   III. Nationhood, Sectionalism, and The Civil War: 1815-1877
      a. Economic and social change (e.g., growth of cities, industrialization, transportation)
      b. Jacksonian Democracy (e.g., political parties, expanding the franchise, treatment of Native Americans)
      c. Expansion of Slavery (e.g., Missouri Compromise, plantation, economy, abolitionists)
      d. The Civil War
   IV. Territorial Expansion, the Rise of Modern America, and World War I: 1877-1920
      a. Territorial Expansion
      b. The Rise of Modern America
      c. The First World War
      d. Women's Vote - Nineteenth Amendment (e.g., early advocates: Susan B. Anthony, Elizabeth Cady Stanton, Seneca Falls)
   V. The Great Depression, the New Deal, and World War II: 1920-1945
      a. The 1920s (e.g., temperance movement and prohibition, inventions, Scopes trial)
      b. Causes and characteristics of the Great Depression (e.g., stock market crash, collapse of economy, Dust Bowl)
      c. Franklin D. Roosevelt and the New Deal (e.g., changes in role of government, gains for labor, agricultural price supports, Social Security)
      d. The Second War
   VI. Post-World War II: 1945 to Present
      a. The Cold War (e.g., containment of communism, beginnings of arms race, Truman Doctrine, Marshall Plan, NATO, fear of communism leading to McCarthyism, Communist expansion in Europe)
      b. Korean Conflict (e.g., UN forces, MacArthur versus Truman)
      c. Post-war prosperity (e.g., demand for consumer goods)
      d. The 1960s
      e. The 1970s

B. NAEP (1987-88): "We worked out a scheme, we agreed on the (Advisory) Committee that one could divide American History into a number of chronological periods and within those there were various topics that were important and we, in the meetings, we discussed what those topics were. We also agreed that there were themes in American Political History and Constitutional History, Economic History, Social History, including the history of society in general and also the history of particular groups, such as women and various racial and ethnic minorities and the history of culture and literature and so on, that all those were areas where there were long term themes that were better treated, that ought to be identified and weren't necessarily covered if you divided American History into periods which tend to become political periods defined by great presidencies and by wars. So we said all that in that (NAEP) booklet and so we had, we produced both a chronology and a list of major thematic topics along the lines of political, economic, social, cultural..." (Interview D).

I. Exploration and Colonization: up to 1763
   a. Geographic Context
   b. The First Americans
   c. European Exploration
   d. Colonial Development
II. The Revolutionary Era, the Constitution, and the New Republic, 1763-1815
   a. Crisis and Independence
   b. The Constitution and the Bill of Rights
   c. Establishing the New Nation
III. Economic and Social Development of the Antebellum Republic, 1790-1861
   a. Economic Expansion
   b. Industrialization
   c. Political Development
   d. Intellectual and Cultural Life in the Republic
   e. The Problem of Slavery
   f. The New West
IV. Crisis of the Union: Origins of the War, the War, and Reconstruction, 1850-1877
   a. "Manifest Destiny" and Expansionism
   b. Emerging Conflict between North and South
   c. The Civil War
   d. Reconstruction and Constitutional Transformation
V. The Rise of Modern America and World War I, 1877-1920
   A. Economic Expansion
   B. Political Movements
   C. Civil Rights and the Constitution
   D. American Overseas Expansion and Empire
   E. World War I
VI. The United States, 1920-1941
   a. The 1920s
   b. The Great Depression
   c. The New Deal
VII. World War II and the Postwar Era, 1931-1968
   a. World War II
   b. The Cold War Era
   c. Political and Constitutional Change
   d. Economic, Social, and Cultural Developments after 1945
VIII. Modern Post-Industrial Era: 1968 to the present
   a. Political Change
   b. International Policies and Forces
   c. Technological and Economic Change
   d. Social and Cultural Change

C. State Test (the secondary curriculum - Grades nine through twelve):
I. Grade Nine (elective courses in History-Social Studies):
   a. Our State in the Twentieth Century
   b. Physical Geography
   c. World Regional Geography
   d. The Humanities
   e. Comparative World Religions
   f. Area Studies: Cultures
   g. Anthropology
   h. Psychology
   i. Sociology
   j. Women in Our History
   k. Ethnic Studies
   l. Law-Related Education
II. Grade Ten - World History and Geography: The Modern World

a. Unresolved Problems of the Modern World
b. Connecting with Past Learnings: The Enlightenment and the Rise of Democratic Ideas
c. The Industrial Revolution
d. The Rise of Imperialism and Colonialism: A Case Study of India
e. World War I and Its Consequences
f. Totalitarianism in the Modern World: Nazi Germany and Stalinist Russia
g. World War II: Its Causes and Consequences
h. Nationalism in the Contemporary World
The Soviet Union and China
The Middle East: Israel and Syria
Sub-Sahara Africa: Ghana and South Africa
Latin America: Mexico and Brazil

III. Grade Eleven - United States History and Geography: Continuity and Change in the Twentieth Century
a. Connecting with Past Learnings: The Nation's Beginnings
b. Connecting with Past Learnings: The United States to 1900
c. The Progressive Era
d. The Jazz Age
e. The Great Depression
f. World War II
g. The Cold War
h. Hemispheric Relationships in the Postwar Era
i. The Civil Rights Movement in the Postwar Era
j. American Society in the Postwar Era
k. The United States in Recent Times

IV. Grade Twelve - Principles of American Democracy (one semester)
a. The Constitution and the Bill of Rights
b. The Courts and the Governmental Process
c. Our Government Today: The Legislative and Executive Branches
d. Federalism: State and Local Government
e. Comparative Governments, with Emphasis on Communism in the World Today
f. Contemporary Issues in the World Today

V. Grade Twelve - Economics (one semester)
a. Fundamental Economic Concepts
b. Comparative Economic Systems
c. Microeconomics
d. Macroeconomics
e. International Economic Concepts

Examples of materials to be used included the following from the state curriculum framework:

The Soviet Union and China: "Students should read excerpts from Nikita Khrushchev's speech of 1956 denouncing the crimes of Joseph Stalin".

Sub-Saharan Africa: Ghana and South Africa: "Reading excerpts from Mark Mathabane's, *Kafir Boy*, will give students insight into growing up as a black under Apartheid."

The Progressive Era: "By reading some of the extraordinary decisions of Justices Louis Brandeis and Oliver Wendell Holmes, students sill understand the continuing tension between the rights of the individual and the power of the government."

The Jazz Age: "A migration of many blacks from the South helped to create the 'Harlem Renaissance', the literacy and artistic flowering of blacks, artists, poets, musicians, and scholars such as W.E.B. DuBois, Langston Hughes, Richard Wright, and Zeale Neale Hurston. Examples of their work should be read."

The Civil Rights Movement in the Postwar Era: "They (students) should understand Dr. King's philosophical and religious dedication to nonviolence

by reading documents such as his 'Letters from a Birmingham Jail.' Well chosen readings that should heighten students' sensitivity to the issues raised in this unit, such as *The Autobiography of Malcolm X,* Lerone Bennett's *Before the Mayflower: A History of Black America,* Ralph Ellison's *Invisible Man,* Richard Wright's *Native Son,* Lorraine Hansberry's *A Raisin in the Sun.*"
The Courts and the Governmental Process: "...critical reading of the *Yick Wo* decision should serve to remind students that racial discrimination affected not only blacks but other groups as well, including Asians and Hispanics."
Our Government Today: The Legislative and Executive Branches: "Through selected case studies, students can analyze presidential campaigns, the handling of international crises, and the scope and limits of presidential power. Examples might include the Steel Crisis, the Cuban Missile Crisis, or the Iran Hostage Crisis. Students should explore the process of presidential decisionmaking through role play, simulation, and interactive learning."

D. Commercial Test

"Because we believed that there ought to be a lot more integration in this field than sometimes exists, we tried not to isolate things too much. But some of the themes, although they came under something such as history, we did not write items that would only apply to history, we did not write items that would only apply to history but which would relate history to other factors—political science, sociology, geography" (Interview F).

Nine-Ten Level: "Our objectives, general objectives at that level were history, political science, sociology-anthropology, economics, and geography. Then we had two more, one of them is called Interrelated Disciplines and that was a catch-all where we emphasized those that seemed particularly to combine history and geography. But we didn't exclude those combinations in any of the other ones, but perhaps we felt the emphasis was political so we put those in" (Interview F).

Geography: "Locational applications that was mainly map reading and interpretation. Regions and physical features. Human and environmental interaction." Economics: "In economics we tried to cover the fundamental concepts of economics. Economic communities. Resources and technologies. Regional energy dependency and global interdependence." "In history we had time, cognition, and chronology. Significant events and people. These are just a little catch-all phrase. Patterns of civilizations. In political science we tried to cover governmental structure, democratic process, rights, responsibilities. Regional and global issues. In sociology anthropology, individual and group roles.

Cultures. Belief systems. Multicultural perspectives. Issues and attitudes. And then under the objective called Interrelated Disciplines, that's where we dumped things according to their disciplines. If it was something that seemed to strongly link and emphasize the link between geography and history, we'd put it in that category. But there were items which, as I said, emphasized the relationship rather being based on factors from different disciplines, actually emphasized the relationship. Then the last one we applied to Social Studies. We more or less grouped those into assessing information, concepts, and using the information or the context" (Interview F).