Review of NAEP Materials

Submitted to the

Department of Education

for

Evaluation of the

National Assessment of Educational Progress

Prepared by:

Chad W. Buckendahl, Ph.D.
Susan L. Davis, M.A.
Teresa J. Eckhout, M.A.

Buros Institute for Assessment Consultation and Outreach
A Division of the
Oscar and Luella Buros Center for Testing at the
University of Nebraska – Lincoln

Developed in collaboration with:

University of Massachusetts / Center for Educational Assessment

and

The University of Georgia

February 2, 2005

Overview of Studies Reviewed by Topic
Overview of NAEP
The Nation's Report Card: Evolution and Perspectives
Overseeing the Nation's Report Card: The Creation and Evolution of the National Assessment Governing Board (NAGB)
Grading the Nation's Report Card: Evaluating NAEP and Transforming the Assessment Educational Progress
An Agenda for NAEP validity research
The NAEP 1998 Technical Report
Study 1: Conducting an Audit of the NAEP Assessment Programs
NAEP Quality Assurance checks of the 2002 Reading Assessment Results for Delaware
Management and Technical Review of the National Assessment of Educational Progress
Including Special-Needs Students in the NAEP 1998 Reading Assessment: Part II, Results for Students with Disabilities and Limited-English-Proficient Students
The Effects of Accommodations on the Assessment of LEP students in NAEP
The Validity of Oral Accommodation in Testing
Impact of Selected Background Variables on Students' NAEP Math Performance
Optimizing State NAEP: Issues and Possible Improvements
An Investigation of Why Students Do Not Respond to Questions
The Measurement of Instructional Background Indicators: Cognitive Laboratory Investigations of the Responses of Fourth and Eight Grade Students and Teachers to Questionnaire Items.
Evaluation of Bias Correction Methods for "Worst-case" Selective Non-participation in NAEP
Federal Sample Sizes for Confirmation of State Tests in the No Child Left Behind Act
The Effects of Finite Sampling on State Assessment Sample Requirements
Analysis of NAEP Combined National and State Samples
Statistical Power Analysis and Empirical Results for NAEP Combined National and State Samples
Study 2: Evaluating Population Invariance in NAEP Assessments
Proceedings of Achievement Levels Workshop
Growth in School Revisited: Achievement Gains from the Fourth to the Eighth Grade
The Impact of Item Treatments on NAEP Reporting Scale Scores
A Study of Equating in NAEP
Contributions of Background Questions to Improving the Precision of NAEP Results

Study 3: Evaluating the Alignment of NAEP Tests to Selected State Frameworks
Using the National Assessment of Educational Progress to Confirm State Test Results
National Longitudinal School-Level State Assessment Score Database: Analyses of the 2000/2001 School-Year Scores
Student Performance Standards on the National Assessment of Educational Progress: Affirmation and Improvements
A Content Comparison of the NAEP and PIRLS Fourth-Grade Reading Assessment
A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMMS-R), and the Programme for International Student Assessment (PISA)
Study 4: Research on Validity of NAEP Achievement Levels
Trends in Large-Scale Testing Outside the United States
A Response to 'Setting Reasonable and Useful Performance Standards' in the National Academy of Sciences' Grading the Nation's Report Card
Proceedings of Achievement Levels Workshop
Student Performance Standards on the National Assessment of Educational Progress: Affirmation and Improvements
A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMMS-R), and the Programme for International Student Assessment (PISA)
Raising Achievement and Reducing Gaps: Reporting Progress Toward Goals for Academic Achievement in Mathematics
Study 5: Evaluating the Impact of DIF on NAEP
The Impact of Item Treatments on NAEP Reporting Scale Scores
Study 6: Research on the Utility of NAEP Reports
Trends in Large-Scale Testing Outside the United States
Improving NAEP for Research and Policymaking
Reporting the Results of the National Assessment of Educational Progress
Differences in the Gender Gap: Comparisons Across Racial/Ethnic Groups in Education and Work
Perspectives on Background Questions in the National Assessment of Educational Progress
NAEP Background Questions: What Can We Learn from NAEP About the Effect of Schools and Teachers on Student Achievement?
The Measurement of Instructional Background Indicators: Cognitive Laboratory Investigations of the Responses of Fourth and Eight Grade Students and Teachers to Questionnaire Items.

The National Assessment of Educational Progress (NAEP) has been used for decades as a measure of student learning in the United States. During that time, numerous research studies and extensive evaluation work has been undertaken to examine the design, development, implementation, analysis, and reporting involved with the NAEP assessment program. Notable members of the psychometric community, testing contractors, and government agencies have conducted much of this work. In turn, the findings from this body of work have contributed to the redesign and modification of NAEP for improving the psychometric quality of the assessment program.

The purpose of this review was to establish an understanding of the prior research related to the different aspects of the NAEP assessment program, specifically those highlighted as important areas in the current evaluation design. Given this purpose, documents for review were carefully selected based on several criteria. First, the proposed studies reflect areas in which additional evaluation work may be needed. Prior work was included in the review if it were relevant to one of the proposed studies included in the negotiated design. Second, the evaluation team considered the relevance of each piece of work to the current NAEP assessments. Over the years, the design for NAEP has changed significantly which makes some documents more applicable than others. Third, documents could only be included if they were available to the evaluation team. This precluded review of some of the working papers and technical reports that are not yet available. Documents were obtained through searches of the Internet, library databases, archives of organizations, contractors within the NAEP system, and recommendations from ED staff members.

Based on these criteria, a total of 37 documents were selected for review. This group of documents consists of journal publications, technical manuals, books, technical reports, and research reports. Documents were assigned to various members of the evaluation team for review based on their role in the proposed evaluation and their specific areas of expertise. Each member of the team was asked to critically read and particular documents and prepare a review of approximately 1 page in length, which identified the relevance of each document to the proposed evaluation design. After review, some documents became apparently more applicable to informing the evaluation than others and provided insight into specific areas of NAEP that need further examination. Many documents that were reviewed appear to be helpful resources and will serve to inform different studies during the evaluation as they provide detailed documentation on particular NAEP processes.

The current evaluation design is the result of an ongoing revision process during which the ED has suggested changes to the evaluation team concerning the inclusion/exclusion of particular studies. The original proposal design was created to include studies that reflected the issues defined by the Department of Education and Legislative requirements. Through the revision and negotiation process, the original proposal design was reduced to a select group of 5 studies (plus 1 optional study) that reflect the priorities of ED. The following is an overview of the documents reviewed, specifically, their relevance to each of the six proposed studies. This overview is followed by the individual document summaries.

Given the immense scope of the NAEP assessment program, we recognize that this review represents only a small fraction of the work that has been conducted on

NAEP. Additional documents may become available that will be used to inform specific aspects of the evaluation design. In addition, other documents and materials will be reviewed as part of the proposed audit process.

Audit of NAEP Assessment Program (Study 1)

A recent publication (Jones & Olkin, 2004) on the history and development of NAEP discussed the potential for NAEP to serve as a monitor for Federal programs. This may be characterized as one plausible future for the NAEP testing program. Given the importance of NAEP as a measure of educational progress in this country and the need for it to inform other testing programs (e.g., NCLB), careful monitoring is needed of NAEP practices from test development to score reporting and use. The audit study will provide an in-depth psychometric review of the NAEP assessment program by examining all major process involved in the development, sampling, administration, analysis, scoring, reporting, and security of the NAEP assessment program within its validity framework. The audit study will focus on NAEP assessments in reading, mathematics, and science; however, the documents reviewed highlight several particular areas of the program that may also be important for NAEP to serve these additional purposes. For example, one document in this review (AIR, 2004b) presents several issues associated with estimating performance gaps (differences in performance between groups) specifically in relation to using NAEP to confirm or dispute group differences reported from state test data.

Organization of NAEP

The NAEP program represents a complex system of policy makers, testing contractors, and government officials. A major component of the audit study will be to

examine each piece of this system individually as well as part of the larger structure. The KMPG (1996) review of NAEP includes useful information on the organizational structure of NAEP. However, this review appears to lack the technical perspective necessary to provide a critical psychometric review of NAEP methodologies and practices. The audit study proposed for this evaluation will analyze the organization and practices of the NAEP assessment program using both test program management and technical perspectives. Other documents that take more of a historical perspective (e.g., Vinovskis, 1998) will also aid in the understanding specific organizations within the NAEP structure (e.g., NAGB).

The review of literature identified several aspects of the NAEP structure that will be important to examine during the audit process. First, one of the major problems with the NAEP assessment is the lag time in reporting results from each administration (e.g., Jones & Olkin, 2004; NRC, 1999). This lengthy process is likely due to the many levels that are incorporated within the NAEP hierarchy (from policy makers to testing contractors). Part of the audit will be to document the structure of the NAEP system and provide a review of the process through which results are generated and reported. Second, in a recent publication regarding the role of NAGB in NAEP, concerns were expressed over the decision making structure within NAEP (Jones & Olkin, 2004). The policy making structure will also be examined as part of the audit study. Third, NAEP results have the potential to serve a number of purposes (AIR, 1997b). Part of the audit process will be to examine the score reporting process and this study would provide results that could be used in collaboration with the utility study (if it is funded) to best inform the use of NAEP scores.

Validity Checks and Internal Accountability

In 2003, as part of an ongoing quality assurance contract, an analysis was conducted by HumRRO of Delaware's NAEP results (NCES, 2003). As part of the audit, it will be important to determine the level of validity checks (both internal and external) similar to this study that are conducted routinely (and in response to problems) to maintain the psychometric integrity of the testing program and NAEP results. Because of the increasing importance of NAEP results, there should be a system of validity checks that require ongoing analyses and reviews of NAEP policies and procedures. In turn, this system should also incorporate a process by which these findings can inform future NAEP policy and practice.

Sampling

A second area that has received much attention in NAEP research is the sampling design of NAEP assessments. The previous evaluation of NAEP (NRC, 1999) highlighted challenges with sampling. Given the requirements for NAEP participation under NCLB, the sampling issues in recent years are likely different than those in the past. For example, with anticipated linking between NAEP and state test data, selective non-participation by particular groups of students may become a greater concern. AIR (2004a) presents an analysis of the bias that could result from selective non-participation at both the school and individual student level and demonstrates methods for correction of this bias. During the audit evaluation of the sampling design in NAEP, it will be important to determine if checks are in place to monitor selective non-participation and how to deal with the effects of such problems on NAEP data.

One of the most debated issues in any sampling program is obtaining an appropriate sample size to accomplish the needs of the testing program while maintaining a cost-efficient design. Not only must this sample size be adequate to represent the population of interest, statistical theory demonstrates the need for a sample large enough to make meaningful comparisons between groups within the sample. In other words, power becomes an issue. The NAEP validity studies panel commissioned a paper to examine how one investigates performance gaps (differences between groups) in terms of which samples to compare, how to compare performances of the selected samples, and the sample size requirements for such comparisons (AIR, 2004b). The purpose of this paper was to inform sampling of NAEP when comparisons are to be made between groups to confirm state test data. The paper indicated that current NAEP sampling was not adequate to completely confirm or refute group differences in state test data.

Currently, NAEP has reported using combined state and national samples. Substantial work appears to exist supporting the use of these combined samples (e.g., ETS, 2003b, 2003d). Specifically, this work has demonstrated that the combined samples provide increased power and precision without sacrificing substantial efficiency.

Finally, much sampling research has focused on inclusion policies and accommodation allowances for students with disabilities (SD) and Limited English Proficiency (LEP). The NAEP validity studies panel (AIR, 2002a) rated studies in this area as “High” on importance for future research, since requirements for NAEP participation have changed with the NCLB legislation. The sampling policy for students within SD and LEP populations should be considered in relation to the current policies for accommodations in NAEP assessments.

Accommodations

Several student populations within the NAEP sampling framework often require or request accommodations when testing. Much research has centered on the feasibility of using different accommodations and how these affect test performance and the comparability of scores between these students and the rest of the population. Certain accommodations have been shown to reduce construct-irrelevant variance in NAEP scores (AIR, 2002c). With respect to performance by SD and LEP students, the effects of accommodations appear to be mixed. For 8th grade math assessments, Abedi, Lord, and Hofstetter (2001) found that differences in performance between LEP and non-LEP students was not reduced when the English in the test was simplified. However, on a science assessment, Abedi, Lord, Kim, and Miyoshi (2001) determined that a customized dictionary significantly aided in the performance of LEP students.

Often concerns are voiced about the effect of accommodations on the meaning of the assessment. Differential item functioning (DIF) procedures can be used to determine if the construct being assessed changes when accommodations are allowed. ETS (2004) examined this issue for 4th and 8th grade reading assessments and determined that only a minimal number of items exhibited DIF between accommodated and non-accommodated students. Findings such as these suggest that results for SD and LEP students who received accommodations can be included. However, they also suggest the need for research on the utility of accommodations. Through this process, NAEP administrators can identify which accommodations would best serve a particular population of students for a particular content area. The use of accommodations will be explored in the audit and suggested topics of investigation include:

1. Does the accommodation match the construct being tested?
2. Does the accommodation change the validity of the scores being reported?
3. Is the accommodation appropriate for the student(s) who are being targeted (i.e., LEP, blind, hearing impaired, etc)?
4. Has the accommodation been field tested and found to be significantly necessary?

Scoring and score reporting

Finally, it will be important for the audit process to include review of the NAEP system of scoring and score reporting. Given that there may be no incentive for students to perform well on NAEP assessments, motivation becomes an issue when NAEP scores are calculated and reported. Work in this area (e.g., AIR, 1999) suggests that students omit items for various reasons, not just because they find them too difficult to answer. Therefore, missing data from assessments such as NAEP must be documented and handled very carefully.

Evaluating Population Invariance in NAEP Assessments (Study 2)

Work concerning group differences seems to maintain importance in NAEP assessments. The NAEP validity studies panel (AIR, 2002a) labeled studies on population bias as "Highly Important." This report and other studies included in this review highlighted the need for studies concerning group differences and potential test score bias. Many of the reviewed studies will provide important references for the population invariance study depending on which populations are to be included in the analysis plan. For example, references within this body of work were identified that address issues concerning difference between states (AIR, 1997b), gender, ethnicity

(ETS, 2003a), and over time (NEGP, 2002). In addition, this research provides several technical references concerning the estimation of plausible values for conditioning variables (Reckase, 2002), the sensitivity of results to background variables (AIR, 1997a), bias evident during the NAEP equating process (1997a), and the estimation of person and item parameters (ETS, 2003c). Overall, this work indicated the importance of considering group membership when examining NAEP performance and underscores the need for particular studies on topics such as population invariance to determine if these comparisons are psychometrically valid.

Evaluating the Alignment of NAEP Tests to Selected State Frameworks (Study 3)

One clear theme in many of these documents is the need for NAEP results to serve NCLB purposes (e.g., AIR, 2004). Specifically, policy makers and stakeholders want NAEP data used to confirm state testing results. Before this can happen, numerous steps must be taken to ensure that the comparisons will be meaningful and accurate. The NAEP validity studies panel (AIR, 2002a) categorized the alignment of state content standards to NAEP assessments as "essential", giving it the highest categorization in their classification system. The alignment study will address this issue by exploring the relationship between NAEP assessments and state content frameworks.

While no published studies directly comparing NAEP content to state tests or frameworks could be located, one study (Bourque & Byrd, 2000) found similarities between NAEP achievement levels and state achievement levels. In addition, several studies provide guidelines for comparing content across two different sates (e.g., NCES, 2003a; Nohara, 2001) and using NAEP to match state results (NAGB, 2002).

Research on the Validity of NAEP Achievement Levels (Study 4)

With the increased importance of NAEP results, it is of the utmost importance that the achievement levels on which NAEP scores are reported provide accurate representations of student knowledge. The need to carefully examine the current standard setting methodologies has been identified in the literature (Hambleton et al., 2000). The current evaluation design places a high level of importance on the study proposed to examine the validity of the NAEP achievement levels. Despite the existence of work examining achievement levels (e.g., Borque, 2000), the implementation of new standard setting procedures coupled with the importance of NAEP results suggest that the achievement levels study will be an important aspect of the evaluation.

Several documents that were reviewed will serve as important resources for the background of the standard setting processes used in NAEP (e.g., Bourque & Byrd, 2000; Vinovskis, 1998). In addition, two studies included in this review provide descriptions of how NAEP content compares to the content included in other assessments, specifically that of PISA and TIMSS (NCES, 2003a) and PIRLS (Nohara, 2001). These findings are particularly relevant to the proposed aspect of the achievement levels study where scores will be evaluated against other large testing programs, however, the international comparisons may be beyond the scope of the study depending on degree to which there is overlap among items used on NAEP tests and on these other assessments. Finally, one document takes a closer look at some specific performance based on achievement levels to determine how group performance compared (NEGP, 2002).

Research on the Utility of NAEP Reports (Study 5 – Optional pending additional funding)

The utility study is included within the proposal; however, it will not be conducted unless additional funding is secured. This review noted several places where

such a study is necessary given the importance of multiple constituencies being able to understand NAEP reports and interpret the information correctly. For example, the previous evaluation of NAEP (NRC, 1999) suggested criteria for NAEP reports. Grissmer (2002) demonstrated the perceived limitations of NAEP score reports and made suggestions for improvement. Bourque (2000) presented guidelines for creation of score reports to make them more understandable. If funded, the utility study would be used to determine if these criteria could be met. In addition, several studies discussed the potential utility of information provided from background questions in NAEP reports (e.g., Barton, 2002; Podgursky, 2002). Included in the utility of NAEP reports will be an examination of reporting group differences. If the utility study is funded, the final report will be based on a framework described by Jaeger (AIR, 1998a). This framework includes a series of questions regarding the best format and content for NAEP reports.

Evaluating the Impact of DIF on NAEP (Study 9)

Similar to the population invariance study, the exact relevance of each of these documents to the impact of DIF study will be determined after the list of included groups is defined. Several documents indicate the importance of including special needs populations within NAEP reports and the levels of DIF exhibited between SD and non-SD students (e.g., Abedi et al., 2001). In addition, the DIF study will explore trends in DIF over time therefore, resources that detail item parameter estimation will be important (e.g., ETS, 2003c).

DOCUMENT REVIEWS

Title: Impact of Selected Background Variables on Students' NAEP Math Performance

Authors: Abedi, Lord, and Hofstetter (2001)

Purpose: The purpose of this study was to examine the effects of test modifications on NAEP math performance for students whose first language was Spanish.

Findings:

The study included 1394 8th grade students from southern California schools, of whom nearly 2/3 were classified as Limited English Proficient (LEP). LEP students were randomly assigned to take one of three versions of the test: an English modified to make the language simpler (English modified), the original version of the test in English, or the original version of the test in Spanish. Students who were not English Language Learners were not given the Spanish version. NAEP reading scores served as a covariate to control for differences in reading ability. Results indicated significant main and interaction effects for test version. However, inspection of the table of means reveals that the LEP students did no better on the modified English version than on the original English version of the test (means of 11.79 and 11.84, respectively), while students who were not LEP did better on the modified English version than on the original English version (means of 16.71 and 15.26, respectively). Thus LEP students did not appear to be advantaged by taking the version of the test using simplified English, while non-LEP students were. Additionally, LEP students did worse on the Spanish version of the test (mean = 9.16).

Relevance to the current evaluation:

The results of the Abedi et al. study suggest that differences in scores of LEP and non-LEP students are not ameliorated by simplifying the English reading level of the items, which implies that differences may not be due to language problems. Although their study is not a DIF study, and they don't use this term, one possible conclusion that could be drawn from their results is that any DIF found on the NAEP 8th grade math test may not be language-related, at least for Spanish-speaking students similar to those in the study. Thus, other sources of DIF should be considered. In addition, the Audit study (study 1) will examine the use of accommodations in NAEP assessments. This study suggests that selection of accommodations to be made available could be informed by research on the effectiveness of accommodations.

Title: The Effects of Accommodations on the Assessment of LEP Students in NAEP

Authors: Abedi, Lord, Kim, and Miyoshi (2001)

Purpose: The focus of this study is on two forms of an accommodation targeting the non-technical vocabulary of test items, which have been identified as the source of difficulty for LEP (Limited English Proficiency) students. Both LEP and non-LEP students were tested under accommodated and non-accommodated settings. Accommodation strategies chosen for this study were considered practical for use in large-scale assessments.

A NAEP science assessment consisting of twenty items was administered in three forms. One contained the original items with no accommodations and two with accommodations, which focused on potentially difficult English vocabulary. One of the accommodated forms contained a customized English language dictionary at the end of the test booklet. The other accommodated form contained English "glosses" (an individual definition or paraphrase) and Spanish translations in the margins of the test booklet.

Findings: LEP students performed lower than non-LEP students. The difference between the two groups is significant ($t=6.83$, $df=417$, $p=.000$). LEP students performed substantially higher under accommodated conditions than under the standard condition. For the students under the customized dictionary the mean score was 10.18 ($SD=5.26$, $n=55$); under the glossary condition the mean was 8.51 ($SD=4.72$, $n=70$); and under the standard condition the mean was 8.36 ($SD=4.40$, $n=58$). Analysis of variance indicated the difference between means for LEP and non-LEP students under the three conditions was significant ($F=3.08$, $df=2,180$, $p=.048$).

The scores had no significant effect on the scores of the non-LEP students. For non-LEP students the mean score for the dictionary was 11.37 ($SD=3.79$, $n=82$); for the glossary it was 11.96 ($SD=3.86$, $n=75$); and for the standard condition the mean was 11.71 ($SD=3.40$, $n=79$).

This study showed that the customized dictionary significantly enabled LEP students to perform better on the science assessment, and that the accommodation strategies used did not impact the construct and validity of the assessment. The authors point out that these results are particularly encouraging given the ease with which the accommodation is administered.

Additional findings include analyses on background variables of the test takers participating in this study. A significant difference was found between the performance of students who speak only English in the home and those who speak another language other than English in the home. Students who speak a language other than English performed significantly lower than the other group.

Relevance to the current evaluation:

The proposed evaluation addresses the use of accommodations within the scope of the audit study. The use of accommodations should be carefully evaluated to determine that the construct of the assessment and the validity of the scores are not being

compromised. Possible additional questions to add to the audit study to address these issues might include:

1. Does the accommodation match the construct being tested?
2. Does the accommodation change the validity of the scores being reported?
3. Is the accommodation appropriate for the student(s) who are being targeted (i.e LEP, blind, hearing impaired, etc)?
4. Has the accommodation been field tested and found to be significantly necessary?

Title: Evaluation of Bias Correction Methods for 'Worse-case' Selective Non-participation in NAEP.

Organization: American Institutes for Research (2004a)

Author: McLaughlin, Gallagher, Stancavage

Purpose:

NCLB legislation has resulted in a requirement of states and schools to participate in the NAEP assessment program. However, participation by individual students remains voluntary. Therefore, NAEP administrators must consider the effect that non-participation of particular groups of students will have on NAEP results. Specifically, states and/or schools may inevitably encourage non-participation by low performing students to inflate their overall performance scores. In addition, non Title-I schools could decide not to participate in NAEP, which could significantly change the sample for the entire state. Therefore, the purpose of this research was to determine the extent that NAEP results would be biased if selective non-participation were to occur (at either the individual level or the school level) and how this bias could be corrected through various statistical methods.

Findings:

Bias was simulated at both the school and individual level. Analyses indicated that state means for NAEP were biased upwards approximately 2.9 NAEP scale points when the lower 10 percent of schools were excluded from the analyses. At the individual level, state NAEP scores were inflated approximately 5.1 score points when the lowest performing 10 percent of students were simulated as "non-participating". Three methods for correcting bias were tested. Methods such as linear equating appeared to reduce the level of bias by approximately half; other methods required state test data to eliminate bias, which is not always available.

Relevance to current design:

This paper raises some important concerns about NAEP participation under NCLB policy. The authors note specifically that as more work is done to link NAEP to state testing, problems with selective non-participation may become more prevalent; corrections methods aided in dealing with the resultant bias. The sampling component of the audit study will be an important aspect of the evaluation process, as issues like this will continue to arise. Within the sampling plan for NAEP, and even as part of the larger policy of NAEP, it will be important to see what is being done to monitor selective non-participation and prevent problems of this nature.

Title: Federal Sample Sizes for Confirmation of State Tests in the No Child Left Behind Act

Organization: American Institutes for Research (2004b)

Author: Mosquin and Chromy

Purpose:

Within educational assessment, it is easy to find reports of differences in student performance based on group memberships (termed in this paper "performance gaps"). The purpose of this work was to define exactly what a gap in performance is and how to identify these gaps using performance on NAEP assessments. Specifically, performance was examined for groups of students based on mean NAEP scores, differences in percentages of students classified as either below or above the "Basic" level of performance, and as either below or above the "Proficient" level of performance. Discussion about performance gaps is particularly important given the requirement in NCLB legislation for states to reduce performance gaps for disadvantaged students.

Findings:

In terms of defining a gap, the authors describe two possible methods: comparing the disadvantaged group to the advantaged group from the current year, and comparing the disadvantaged group to the advantage group from a baseline year. They present both methods and the advantages and disadvantages to each. Similarly, they present the advantages and disadvantages to using each method of identifying gaps (i.e., based on mean scores, classification below or above "Basic", and below or above "Proficient"). Finally, sample size requirements are discussed for detecting performance gaps. The specific requirements will vary by state, how a performance gap is being defined, and how gaps are being identified.

Relevance to the current evaluation:

The NAEP validity studies panel commissioned this paper to examine the plausibility of using NAEP data to confirm state testing results reported under the NCLB requirements. The findings of this review suggest that in its current design, NAEP results could be used to supplement state assessment results. However, NAEP performance gaps findings could not completely confirm those reported from state assessment data due to differences in sample size. To be able to make direct confirmation of state findings, NAEP data would possibly have to be paired with state assessment data on the individual level. In addition, the authors suggest that NAEP results could be used to confirm state's claim of adequate yearly progress.

One theme of the current evaluation is considering how NAEP could be used to somehow address NCLB requirements and this is reflected in the choice of studies to be included in the overall evaluation plan. As this paper indicated, identifying performance gaps is important. While this evaluation design will not directly measure performance gaps, two studies (population invariance and impact of DIF) were designed to examine differences in group performance as is related to the psychometric properties of the test. In addition, the alignment study will be some of the beginning work on comparing NAEP

assessments to state tests/content frameworks. This link must first be established before direct comparisons between NAEP results and state results can be made.

Title: An Agenda for NAEP Validity Research

Organization: American Institutes for Research (2002a)

Author: Stancavage, Beaton, Behuniak, Bock, Bohrnstedt, and Champagne

Purpose: Set the validity research agenda for NAEP in light of implications for NAEP of the No Child Left Behind Legislation.

Findings:

Six main areas for validity research were identified:

Subject domain: What is being measured?
Subject domain: How is it being measured?
Validity Issues:
Representing Populations
Issues and Recommendations on NAEP Data Analyses
Validity and Utility Issues In NAEP Reporting and Data Releases
Estimating Trends from NAEP Scores.

Panel members divided into teams to develop research studies in these six areas. Reports and presentations of these research studies were made to the full panel. Panelists then rated the studies on a 5-point scale from "Essential" to "Not Needed".

Relevance to the current evaluation:

1. Among the "Essential" studies were ones to address the alignment of state content frameworks to the NAEP assessment.
2. Studies related to population bias and accommodations were rated in the "Highly Important" range
3. Studies pertaining to audience interpretations of NAEP reports were rated as "Not High"
4. Other studies rated as Essential included
   a. Meaning of "confirming state's results"
   b. Limits on NAEP's capacity to evaluate state's results
   c. Bridge studies (to maintain trend information over assessment changes)
5. Studies related to sampling issues, particularly the representation of students with disabilities and LEP were rated as "High".
6. Strategies for streamlining the analytical demands, in order to shorten turn around but maintain quality control and precision of results were also addressed.

Title: National Longitudinal School-Level State Assessment Score Database: Analyses of 2000/2001 School-Year Scores

Organization: American Institutes for Research (2002b)

Authors: McLaughlin, Bandeira de Mello, Cole, Blankenship, Hikawa, Farr, and Gonzalez

Purpose: In this paper, the authors describe the National Longitudinal School-Level State Assessment Score Database, which is a database of test scores for about 80,000 public schools from across the U.S. (49 states, DC, and Puerto Rico are represented). The report compares test results *within* each state across Title I, school-wide Title I, and other schools. The purpose of the paper is to illustrate the strengths and limitations of this database for "answering questions about achievement in schools receiving federal programmatic funding" (p. v).

Findings:

This comprehensive report is over 150 pages, with about 100 pages devoted to informative appendices. The authors concluded that reading and math scores in Title I schools are about one-half standard deviations lower than other schools, with the effects greatest at the elementary level. They also conclude that poverty has a particularly strong negative effect on achievement. However, after statistically controlling for poverty, a trend for positive gains for African American students in Title I and School-wide Title I programs emerged, and was statistically significant in Florida and North Carolina. In general, gains across the two school years studied (1999/2000 and 2000/2001) were greater than losses for the majority of states. Many states did not adequately break down their results by subgroup (e.g., minority groups). Achievement gaps between students with and without disabilities tended to be very large (up to two standard deviations), but were significantly smaller in Title I and School-wide Title I schools. The authors conclude that this database may be useful for comparing state profiles of educational achievement and illustrating how differences in achievement are related to various factors.

Relevance to the current evaluation:

Unclear. The results of the study are a bit dated, but the database itself could be useful in the alignment study for evaluating state achievement trends reported by state testing programs versus state achievement trends based on NAEP reports.

Title: The Validity of Oral Accommodation in Testing

Organization: American Institutes for Research (2002c)

Author: Timothy J. Weston

Purpose:

The purpose of this study was to validate the use of an oral presentation of mathematics test items as a testing accommodation for disabled students. Specifically, the author tested the ability of the accommodation (oral presentation of test items) to reduce the construct irrelevant test score variance. Having a lower construct irrelevant test score variance would mean that the accommodation does not change the test score variance between disabled and non-disabled students; the accommodation has little to do with the content area being assessed.

It was hypothesized that the accommodation of the oral presentation of mathematics test items would boost disabled students scores but not the scores of non-disabled students and reduce construct irrelevant variance due to reading. Also, the accommodated test scores would be more in agreement with teachers' ratings of student ability in mathematics than non-accommodated test scores.

Findings:

The results showed that all students' test scores were significantly higher in the accommodated condition with learning disabled students gaining significantly more than non-disabled students. The role of reading in the effects of the accommodation showed that as reading level increased, the gain from the accommodation decreased for learning disabled students and increased for non-disabled students. Students at lower reading levels (disabled and non-disabled) gained significantly more from the accommodation than students at higher reading levels.

Learning disabled students also gained more on all word problems than on calculation problems in the accommodated condition. Non-disabled students and learning disabled students with higher reading abilities had varying scores on word problems in the accommodated condition, meaning that the accommodation's benefit was item specific. Both groups of students benefited equally from the accommodation on calculation problems.

Teachers' ratings of mathematics ability were more in agreement with the accommodated test scores than the non-accommodated test scores.

The author posed a possible construct irrelevant difficulty caused by the accommodation as impatience exhibited by non-disabled students because of the time needed to finish reading items aloud. Possible construct irrelevant easiness caused by the accommodation was suggested as teachers unconsciously giving away an answer by emphasizing the correct one through voice inflection or body language or paraphrasing questions so the answer is obvious. Additionally, another source of construct irrelevant variance may be due to the fact that the teacher paced the test by reading items aloud; thereby helping students stay on task.

Relevance to the current evaluation:

Implications for NAEP as suggested by the author are to provide disabled students with an accommodation of oral presentation of test items to reflect their true ability in mathematics, as indicated by teacher ratings. This accommodation would be beneficial for disabled students as most NAEP items are too difficult for them and not covered in the mathematics curricula for these students. Studies such as this make it imperative that the audit study addresses the issues of accommodations in NAEP assessment.

Title: An Investigation of why students do not respond to questions

Organization: American Institutes for Research (1999)

Authors: Jakwerth, Stancavage, and Reed

Purpose: The purpose of this study was to examine reasons why students fail to respond to particular questions on NAEP assessments. This investigation focused on 8th grade classes where reading and civics NAEP assessments were being administered. Samples of students were interviewed about their reasons for not answering questions. The authors felt that if they determined why students were omitting questions and the potential for them to answer these questions correctly, they could provide useful information for handling missing data in NAEP assessments.

Findings: Out of a sample of 84 students, 65 had omitted at least one question. The range of questions omitted (not counting questions students failed to reach) was 1-12 (average=2.8) and 1-6 (average=2.6) for the reading and civics tests, respectively. Students tended to omit short or extended answer questions more frequently than multiple choice questions. Through the interview process, the researchers identified three main reasons why students skipped questions. First, some students claimed that they did not understand the question, did not know the meaning of one or more words in the question, or did not know the answer to the question. Second, other students claimed they inadvertently skipped the question or did not see it in the test booklet. Third, several students indicated that they were unmotivated or simply did not care about the test. Finally, some students admitted that they skipped items with the plan to return to them later but ran out of time before they could do so. The researchers reported that approximately two-thirds of the students could correctly answer at least one of the questions they skipped.

The authors made several recommendations for future NAEP activities to decrease item omission rates based on the findings of their research. First, during instrument development, items writers should be careful to use appropriate vocabulary and sentence structure/length, and also to make the items relevant to the testing population. For test administration, attention should be given to testing factors such as time, item format, and testing environment.

Relevance to the current evaluation:

The recommendations noted apply to several aspects of the testing process, including item development and test administration. Student motivation is an important factor considering that students do not receive any benefit for their performance on NAEP assessments. A lack of motivation could threaten the validity of the NAEP assessment results so attention to details such as these during the test design process is very important. Part of the audit study incorporates an examination of the scoring and score reporting process. NAEP contractors should be mindful of issues such as item omission and have procedures in place to account for missing data and ensure that it is handled properly. In addition NAEP scores should be reported under the caveat that motivation may be an issue with these scores as indicators of student knowledge.

Title: Reporting the Results of the National Assessment of Educational Progress

Organization: American Institutes for Research (1998a)

Author: Jaeger

Purpose: In this study, Jaeger documents the numerous and varied audiences for NAEP reports and suggests ways in which NAEP reports may be better communicated to specific audiences.

Findings:

The report includes a history of NAEP reporting, which began as item-level reporting. It also describes current reporting as well as proposed alternatives such as "market basket" reporting. It summarizes previous research on the utility of score reports (e.g., Hambleton & Slater, 1995; Hawkins, 1995; Jaeger, 1996). Based on this review, Jaeger raises three questions that could "undergrid a program of inquiry on reporting and disseminating NAEP results" (p. 19). These questions are:

1. In what form should NAEP results be reported?
2. How should NAEP results be displayed?
3. How should NAEP results be disseminated?

He also posits five questions related to the first area of inquiry:

1a. What do various NAEP audiences find to be of interest?
1b. What do various NAEP audiences find to be useful?
1c. What do various NAEP audiences understand?
1d. What can various NAEP audiences validity interpret?
1e. Among alternatives, what do various NAEP audiences prefer? (p. 19)

Jaeger presents a long list of current and potential ways to disseminate NAEP results including web-based, television, and radio approaches. He also outlines the types of information to be conveyed to specific audiences. This outline is provided in the form of a two-by-two table for each audience. The tables are organized by research question and type of data reported.

Jaeger concludes by suggesting several studies on NAEP reporting including (a) research on the influence of press releases, (b) research on improving the understanding of NAEP reports, and (c) research on the dissemination of NAEP results to state education personnel.

Relevance to the current evaluation:

Although the report was written over six years ago, it was written by one of the greatest psychometricians of our generation and it is still very relevant. Should the proposed utility of NAEP score reports study be funded, this report would provide a useful framework for organizing that report. It is also relevant to out achievement levels studies.

Title: The Effects of Finite Sampling Corrections on State Assessment Sample Requirements

Organization: American Institutes for Research (1998b)

Author: Chromy

Purpose:

The sampling design for NAEP at the time of this research required a particular number of schools per state to participate and a particular number of students per school to participate. This often presented problems in obtaining samples that met the required sizes. Rules exist regarding the use of finite population correction factors to estimate variance components. NAEP sampling presents an interesting situation where traditional rules of using finite population correction factors may not apply. The purpose of this study was to examine how finite population correction factors could be applied to between school variance components to examine the effects of using samples of different sizes on precision.

Findings:

Ideally, smaller samples for NAEP assessments would allow for lower administration costs and time. The results indicated that precision could be maintained with smaller samples per state depending on the state demographics. To reduce samples sizes for states, initial work would have to be done to determine appropriate sample sizes and sample representation for each individual state.

Relevance to current design:

This paper provides some interesting perspectives on the NAEP sampling situation. Unfortunately, with NCLB legislation, NAEP sampling and participation requirements have changed to require participation by particular schools. While these findings can still be applied to sampling designs today, the flexibility of sampling within states may be more restrictive than conceptualized in this paper. Part of the audit study will be to evaluate the sampling component of the NAEP assessment. One particular aspect of the sampling evaluation will be to determine if NAEP sampling plans provide sufficient samples and also, if the sampling is done in an efficient manner.

Title: A Study of Equating in NAEP

Organization: American Institutes for Research (1997a)

Authors: Hedges and Vevea

Purpose:

This is a simulation study that was designed to explore the differences between equating procedures for tests of complex designs, such as NAEP.

Findings:

Although it was desirable to make the data as realistic as possible, and measures were taken to do so, the simulated data consisted only of dichotomous items, since software was not available for the mixed format tests. Data were generated using the operational item parameters, and ability distributions were also based on random samples of examinees from the 1992 and 1994 reading and math assessments. Three linking designs were investigated: fixed common item parameter, mean-sigma, and concurrent calibration. A lot of detail is provided regarding the generation of the data and the simulation of the test administration procedure. It appears to mirror the process of NAEP very closely, with the exception of the lack of polytomous items. The results of the simulation indicate the CC and mean-sigma methods perform well, even when each examinee sees very few items, there is multidimensionality, and the ability distributions of the two samples are different, while the FCIP method introduces substantially more bias. However, for all methods, although the estimation of means is relatively unbiased, there was substantial bias at the extremes of the distribution. Additionally, the authors caution that their results were very sensitive to the choice of background variables used in the conditioning.

Relevance to the current evaluation:

This study is very thorough and provides insight into the NAEP estimation and linking process through the description of the simulation study. Its generalizability to operational NAEP is questionable, given that is a simulation study, and neglects constructed response items, which would likely affect the results of the equating. However, given the sensitivity of the results to the background variables, it would be advisable to consider the role of the background variables in the completion of the population invariance study. Further, the bias in the extremes of the distributions has implications for looking at trend lines, even if there is no implication for estimating mean differences. Thus, this study is very relevant to our population invariance studies.

Title: Optimizing State NAEP: Issues and Possible Improvements

Organization: American Institutes for Research (1997b)

Author: Mullis

Purpose: The purpose of the study was to examine ways to optimize the NAEP state assessments given the limited federal and state resources. The study discussed ways to make state NAEP more efficient while retaining its goal of allowing states to compare themselves with national trends and with each other. Topics discussed included: efficiency and sample sizes, schedule of state assessments, content coverage and background questionnaires, linking state assessments to NAEP, and promoting use of NAEP data.

Findings:
Recommendations for optimizing state NAEP included:

- Reducing the burden for states – There is a need for further research on developing more efficient assessment and sampling procedures that will reduce the burden for states. Suggestions to combine national and state samples or to reduce the sample sizes within states were found unfeasible at the time. The major recommendation for reducing the burden for states was to continue to administer state NAEP on an infrequent schedule (every two years) and keep the number of subjects and grades to a reasonable level.

- Establishing a stable assessment schedule –A consistent state NAEP schedule would facilitate participation, provide ongoing trend data, and optimize the use of NAEP in relation to each state's own assessment program. The proposed schedule should be manageable and not exceed the burden currently required by states.

- Promoting the use of state NAEP data – Continued support for state NAEP depends in part on its utility to the states. Suggestions to promote the use of NAEP data included: examining the possibility of linking state assessment and NAEP results, producing more timely and user-friendly reports, and considering shared responsibility between NAEP and participating states.

Relevance to the current evaluation:

This study is probably most relevant to the overall audit of NAEP with respect to meeting the purposes of State NAEP and understanding intent of sampling and recruiting processes. It may also be useful for the invariance study, if invariance over specific states is investigated.

Title: Perspectives on Background Questions in the National Assessment of Educational Progress

Author: Barton (2002)

Purpose:

The purpose of the paper was to provide an argument for the use and importance of background questions on NAEP.

Findings:

Barton described several valuable roles and purposes for using background information on NAEP. First, background information enables the disaggregation of achievement data for subgroups, such as those defined by race and ethnicity and socioeconomic status (SES), so that achievement may be studied for important manifest groups. Second, background questions make it possible to track factors that are related to achievement, which may provide useful information about the potential reason(s) for educational progress. Third, it is possible to track the characteristics of children who are "left behind," which is particularly important considering recent legislation (i.e., No Child Left Behind). Fourth, with background questions, NAEP data may be a useful resource for researchers to explore factors that explain educational achievement. Fifth, NAEP has the potential to provide access to an enormous wealth of information extending beyond achievement to other factors in the school and classroom as well as to the individual student and even into the home. Sixth, by customizing the questionnaire to a state's needs, it is possible to use the information to help education reform. Seventh, certain background factors that are related to achievement may be used to aid estimation of achievement scale scores. And lastly, in order to reap the benefits of background information, it is necessary to invest time and resources into questionnaire development and analysis as well as planning for the use of the data.

Relevance to the current evaluation:

This study would be very useful to the utility of NAEP score reports study, if it becomes funded. It may also be useful in our studies on the achievement levels.

Title: Student Performance Standards on the National Assessment of Educational Progress: Affirmation and Improvements.

Authors: Edited by Bourque and Byrd. Authors Include: Brown, Reckase, Forsyth, Nellhouse, Hambleton, Meara, Popham, Simmons, and Mwalimu (2000)

Purpose: The purpose of this document is to report on the status and process of NAEP by reviewing the policy and practices of the NAGB governing board, public perception of the standards, the commercial and State processes, alternate models for developing achievement levels, and options for modifying the current achievement levels.

Findings:

Study 1: A compilation of the evaluation reports conducted on NAGB achievement levels (AL) since 1990. It discusses the criticisms and decisions that have been made regarding the AL as well as the research ACT has conducted and implemented throughout the years on standard setting.

Study 2: A detailed description of the standard setting process. The author evaluates and describes many different standard setting methods as well. The author "mildly" endorses the use of the bookmark method, the anchor-based procedure, the generalized examinee-centered method, and the multi-stage aggregation method.

Study 3: Describes the standard setting process of three testing companies (CTB, Harcourt, and Riverside) and compares them to NAEP. The results indicated that there were large differences between the percentages of students in each performance level category.

Study 4: The author surveyed States to determine the similarities to NAEP's AL categories and standard setting procedures. Many States (23) have similar AL categories to NAEP and use one of three methods to set standards: modified Angoff, bookmark, or booklet classification. The author also noted that more students were in the Advanced performance level and less students in the lowest performance level category than NAEP.

Study 5: The authors conducted an extensive search of newspaper articles related to NAEP. The authors concluded that more information has been included in the press packages but misinterpretations and causal inferences are still falsely made by the press.

Study 6: The author offers his advice on improvements to the AL that would make it easier to communicate to the public improvements to student performance. They are: add one or more AL's; divide the current AL's into distinguishable, within level reporting categories; make Below Basic a reporting category; re-label existing AL's; lower the scale score ranges associated with one or more AL's.

Study 7: The authors gathered four focus groups to explore their thoughts on the NAEP AL's. Overall, all four groups agreed that NAEP AL's cannot be compared with results from other standardized assessments, supported the release of items, and NAEP data is important to States.

Relevance to the current evaluation: High. These studies will be useful for the achievement level studies and the alignment studies. The report provides a good overview of standard setting activities and research, and outlines some criticisms. Study 4 will be relevant to the alignment study.

Title: Proceedings of Achievement Levels Workshop.

Authors: Edited by Borque. Authors include: Forsythe, van der Linden, Thissen, Billeaud, McLeod, Nelson, Swygert, Hambleton, and Popham. (2000)

Purpose: The purpose of this workshop was to explore the continuity of the different components of NAEP from the achievement level frameworks through score reporting. Each author wrote specifically on a certain topic (i.e., achievement levels, assesembly of test forms, test scoring, score reporting, and standard setting) and concluded with suggestions or comments regarding their findings.

Findings:

Achievement levels: 1) Frameworks are too general to be considered for student scores and frequency in which an examinee could accomplish a task should be included in the frameworks to help standard setters derive a specific performance level on a NAEP scale. 2) Explicit examples should be illustrated in the frameworks since there is a lack of consistency between raters on the categorization of cognitive complexities. 3) Explore the similarity of ALD's across grades since they all should be "grade appropriate" materials.

Assembly of test forms: Two different methods were used to assemble items into item blocks and item booklets (i.e., method 1, items are assigned directly from test pool and method 2 was a derivation of the balanced incomplete block (BIB) used by NAEP). The results indicated that each method is useful for large-scale assessments and ultimately it depends on the test developers' decision to control test attributes (i.e., content, cognitive levels, and # of items in each block).

Test Scoring: This paper introduces the topic of item testlets for items that have more than two response categories. The author suggests that summed scoring (testlets) could nullify the problem of the locally dependent items and enhance the accuracy of the reliability and standard errors of a test. The second article discusses using a "hybridization of summed-scored and response-pattern computation of scaled scores." It compares the Rasch model and 3PL IRT models and concludes that the 3PL/GR scores may be more valid.

Score Reporting: The author noted that many individuals who use the score reports do not fully understand the score reports and the data displays need to be enhanced for easier interpretation.

Standard Setting: Both papers talk about guidelines to improve the standard setting process that can be incorporated to accumulate evidence regarding the inferences of the achievement levels.

Relevance to the current evaluation: Although the studies are a bit dated, they are still relevant resources for understanding perceived limitations of NAEP in terms of achievement level setting, scoring, scaling, and score reporting. This document will be relevant to the achievement levels studies and population invariance study, and to the utility of NAEP score reports study, if it is funded.

Title: Including Special-Needs Students in the NAEP 1998 Reading Assessment Part II

Organization: Educational Testing Service (2004)

Authors: Lutkus in collaboration with Mazzeo, Zhang, and Jerry

Purpose: Since 1996, NAEP has worked to include data from Limited English Proficiency (LEP) students as well as Students with Disabilities (SD) in the assessment sample. As part of this process, NAEP officials were interested in whether allowing accommodations to these students would change the constructs being assessed. In 1998, NAEP reading was administered to a split sample: half of the SD and LEP students were allowed to use accommodations if they normally received them in school, the other half of the sample was not permitted accommodations. The purpose of this work was twofold. First, Differential Item Functioning (DIF) analyses were used to examine the performance of special-needs-students to determine if there was variation in the constructs assessed between accommodated and non-accommodated students. Second, this research examined the performance (on NAEP reading) of SD and LEP students separately using the aggregated national public school and state samples.

Findings: Concerning the first purpose of the research, only a few items exhibited statistically significant DIF for accommodated vs. non-accommodated special-needs students. Therefore, the researchers concluded that accommodations for both students with disabilities and limited English proficiency students could be provided without significantly impacting the results of the aggregated sample for reading.

Concerning the second purpose, students in the SD and LEP groups performed worse on the NAEP exam as compared to those students not classified as SD or LEP. More importantly, the performance of non-accommodated SD students was compared to that of students with disabilities who were permitted accommodations. Results indicated that while there were no significant differences in the performance of these two groups of students at the 4th grade, the 8th grade students without accommodations had higher scores than the group of students who had accommodations.
The results section details the differences within these two populations (students with disabilities & Limited English Proficiency), such as gender, ethnicity, and type of disability.

Relevance to the current evaluation:

This research report raises several important issues that should be addressed by the audit study. First, the inclusion policies described in this report for SD and LEP students are structured so that the final inclusion decision is left to the school IEP team or school administrators. Since the importance place on NAEP has been heightened by NCLB, has the inclusion policy changed? If so, does this new inclusion policy seem appropriate to create standard inclusion across states? Second, the authors conclude that due to the low level of DIF exhibited on NAEP tests concerning SD and LEP students, they see no reason why the data from these populations cannot be included within the larger sample of NAEP reading data. The audit study will examine the current reporting

policies for NAEP data, including the incorporation of SD and LEP student data as well as the sampling policies for these populations.

Title: Growth in School Revisited: Achievement Gains from the Fourth to the Eighth Grade

Organization: Educational Testing Service (2003a)

Author: Coley

Purpose: Combines achievement and growth measures for evaluating student performance on NAEP reading and mathematics tests and compares subgroup performances on state and national levels.

Findings:

- The analysis of average scores on NAEP reading assessment at Grade 4 in 1994 showed that females outperformed males; White and Asian students outperformed Black, Hispanic, and American Indian examinees; nonpublic school attendees outperformed public school attendees; inner city schools' students were outperformed by students in suburbs and large towns. The comparisons of Grade 8 performances in 1998 showed an increase of 50 points while the pattern of the gaps between the subgroups stayed the same. The *extent* of growth among the subgroups was, however, different: Black students showed more improvement than White and Asian students.
- In looking at the average performances of the states on the reading assessment, it was found that Maine and Connecticut were at the top of the list while DC, Hawaii, and California were at the bottom of the list. The analysis of the growth between fourth and eighth grades yielded different rankings of the states: DC, California, and Louisiana were at the top of the list while Maine, Massachusetts, and Minnesota were at the bottom.
- The analysis of average scores on NAEP math assessment at Grade 8 in 2000 showed that males outperformed females; White students outperformed Black, Hispanic, and American Indian students; nonpublic school attendees outperformed public school attendees; inner city schools' students were outperformed by students in suburbs and large towns; students who were deemed eligible for free/reduced lunched were outperformed by those who were not; students from the Southeast were outperformed by students in the other parts of the country. The comparisons of Grade 4 performances to Grade 8 performances showed an increase of 50 points while most subgroup differences were non-significant except that White students demonstrated greater growth relative to Black and Hispanic students. Also, students who were not eligible for free/reduced meals gained significantly more than those who were eligible.
- Examination of average performances at the state level showed that Minnesota, Montana, and Maine were at the top of the list. The analysis of growth between the fourth and the eight graders showed that Montana, Vermont, and Oregon exhibited a higher growth rate than the other states.

Relevance to the current evaluation: The examination of growth in average NAEP reading and math scores compliments the information provided by analyzing the mean

differences of scores within a cohort. This study will be a good reference for the population invariance study.

Title: Analysis of NAEP combined national and state samples

Organization: Educational Testing Service (2003b)

Authors: Qian, Kaplan, and Weng

Purpose: As of 2002, NAEP assessment results will be reported from combined samples including both state and NAEP findings. In preparation for this change, the purpose of this study was to assess the plausibility of merging national and state sample data from NAEP. Specifically, the researchers were interested in examining the following characteristics of the combined sample:

1) How it compared to the national sample as far as overall scores
2) How the groups comparisons from the combined sample compared to those from the national sample
1) The necessity for post-stratification in the combined sample. In the past State NAEP results have been post-stratified whereas National NAEP results have not. To successfully combine the two samples, one would have to explore the feasibility of using non-post-stratified State NAEP samples.

Findings:
The means for combining the sample through a complex system was formulated and executed by Westat.

- Differences between combined sample and national sample were small when comparing mean scale scores and distributions among achievement levels.
- When examining the change in significance testing, it appears that there were more significant differences in comparisons using the combined sample over the national sample. The combined sample provides greater power when identifying differences in proficiency.
- No difference between post stratification and non post-stratification samples (both combined)

In conclusion, the authors determined that combined samples will provide efficient score measures both at the scale score level and achievement categorization level. In addition, the standard errors of the scores will be smaller. The results also suggest that the samples do not need to be post stratified. The standard errors will be slightly larger than for the post-stratified sample, the differences are not likely to be significant. As of 2002 Westat will be responsible for administering both national and state tests and therefore, there will no longer be differences in administration.

Relevance to the current evaluation:
This study was based on data from the 1998 and 2000 administrations of NAEP and supported the use of the combined sample to increase power without sacrificing accuracy. As part of the audit study, the sampling plan will be reviewed including the use of combined samples.

Title: The Impact of Item Treatments on NAEP Reporting Scale Scores

Organization: Educational Testing Service (2003c)

Authors: S-H Tay-Lim, Zhang , Davis, and Tang

Purpose:

The goal of this paper is to determine how much impact the deletion of items from scaling (stringent vs. lenient strategies) has on the estimation of proficiency and to determine some guidelines that guide the deletion of items.

Findings:

A brief description of the calibration of item parameters and person parameters is given. Items are evaluated to see how well they fit the item response data, and a decision is made whether or not to include the item (in say, the equating). A balance is desired between being too lenient or too strict. Current practice relies on a visual inspection to determine whether an item is misfitting or not. Operationally, chi square fit indices are used to roughly detect the misfitting item, but given the issues of power and the correct specification of the degrees of freedom, visual inspection of the empirical and theoretical item response functions (IRFs) is the primary criterion. If an item is deemed misfitting there are several strategies employed to deal with the problem; these are referred to as "treatments." Treatments include: excluding items from the scale, deleting the item from the equating of trend NAEP tests, collapsing categories in the case of polytomous items and combining multiple-choice items into testlets when there is a high degree of dependence. One focus of the study is to investigate the impact of each treatment on the proficiency estimates, as well as to determine when a treatment should be applied. Test information functions are compared for the different treatments. Information functions of the operational treatments are compared to two other strategies for treatments: a more lenient strategy and a more stringent strategy. The results showed the there was a small effect on the information functions whereby the information functions were quite similar between the lenient and moderate (operational) strategies, while there was some loss in information in the stringent case. In terms of proficiency scores, there was almost no difference regardless of the strategy.

Relevance to the current evaluation:

This study presented some interesting information regarding how misfitting items are handled operationally in NAEP. It is essential that in all studies done, particular care be taken to employ the same strategies as are used operationally. The study presented a brief description of how the item/person parameters are estimated, which is particularly relevant to our population invariance studies and to our proposed DIF studies.

Title: Statistical Power Analysis and Empirical Results for NAEP Combined National and State Samples

Organization: Educational Testing Service (2003d)

Author: Qian

Purpose:

Significant discussion has centered on the use of combined NAEP samples representing national and state samples. By using combined samples, the requirements for state testing can be reduced. This paper served to describe the linking process by which state and national samples for NAEP reading assessment are combined and examine the statistical power of the resultant combined samples. This paper was particularly important because starting in 2002 NAEP was to begin reporting results based on the combined sample.

Findings:

The goal of using combined samples would be to produce results with higher precision. This is potentially at the sacrifice of efficiency. This study sought to compare the precision gained to the efficiency lost when using combined samples. The results indicated that confidence could be placed in using combined samples for NAEP reporting. Specifically, the combined samples provided measures of performance as effective as those from the National sample. This finding was replicated across results reported at as mean scores and achievement levels. In addition, the precision of the combined sample was higher than that of the National sample because the standard errors were smaller. Smaller standard errors and more effective samples sizes will likely result in more significant differences between groups in NAEP results. In other words, power to detect group differences will be increased for the larger combined samples.

Relevance to current evaluation:

The author highlights the precision that must be used when combining samples. Specifically, the process of quality control in carefully checking the samples is described. In the current evaluation, one component of the audit will be to address how the scores are calculated and reported including the quality control of this process. While this report clearly suggests that using combined samples will produce results that are effective and precise, ongoing analyses of this nature would ensure that using such samples is continually appropriate.

Title: Differences in the gender gap: Comparisons across racial/ethnic groups in education and work.

Organization: Educational Testing Service (2001)

Author: Coley

Purpose: This study describes a summary of studies designed examine gender differences within racial/ethnic groups**.** It presents a compilation of results that describe the gender differences in a variety of outcomes including NAEP test results. These results and this review are informative about gender differences.

Findings: Overall, the findings suggest little variation in gender differences across ethnic groups. Across racial/ethnic groups, females routinely scored higher than males in NAEP reading and writing. Males scored higher than females on NAEP math examinations at the 4^{th} grade; however, differences at the 8^{th} grade level were only present in one of the samples (i.e., in 1992 but not in 1996). Gender differences varied somewhat by racial/ethnic group for NAEP civics and science.

Relevance to the current evaluation:

It is important to recognize that Coley presents a review of studies of gender differences, not of gender DIF. There is no mention of differential item or test functioning or of item bias in the review. DIF studies are done in an attempt to determine whether the extraneous dimensionality in a test is somehow related to membership in particular groups. In the case of gender DIF, the concern is to determine whether this variability is related to gender. Likewise, the concern for ethnic group or social status group DIF would be to determine whether this extraneous dimensionality is related to either ethnic group or social status group, respectively. Although the gender gap is a serious issue, it is not directly related to the study of differential item functioning nor does Coley tie it to the study of differential item functioning. As previous research has demonstrated, gender DIF on individual items does not necessarily combine additively to cause gender differences.

There is elsewhere research discussing the relation between gender as a cultural factor and DIF. Although Coley does not reference this research, it is not in conflict with the gender DIF literature as summarized by Coley. At this point, however, there are only a couple of studies, most of them preliminary and scattered and, therefore, not providing sufficient evidence to warrant inclusion in the current evaluation.

Title: Improving NAEP for Research and Policymaking

Author: Grissmer (2002)

Purpose: The purpose of the paper was to examine ways to improve NAEP to allow for better research and policy analysis. The paper evaluated the use of alternate data sources that can be used to supplement NAEP background data. The paper also examined ways to strengthen the research potential of NAEP through the selection of background questions and redesigning of the NAEP sample.

Findings:
The author recommends instituting a research panel to advise NAGB in improving NAEP as a research tool. Suggested functions of the panel included: developing an analytical framework for the NAEP background questionnaire, selecting alternate data sources to supplement NAEP data, and examining changes to sampling procedures.

A review of the current NAEP background variables revealed a need for better family characteristic variables and better resource and expenditure information. The author argues that without these variables, inferences made based on other variables currently collected in NAEP may be problematic.

One way to supplement NAEP background variables is to collect data from alternate sources and add them to the NAEP database. Alternate sources discussed included Census data, the Schools and Staffing Surveys, and the Common Core of data. A limitation of using these sources is that they do not provide school level data as NAEP does. Therefore, Grissmer suggests exploring the possibility of redesigning the NAEP sample to allow for district level analysis.

Relevance to the current evaluation:

This study is relevant to the proposed study on the utility of NAEP score reports. If that study becomes funded, this document would be a good source for citing perceived limitations of NAEP reports and suggested improvements.

Title: A response to "Setting reasonable and useful performance standards" in the National Academy of Sciences' Grading the Nation's Report Card

Author: Hambleton, Brennan, Brown, Dodd, Forsyth, Mehrens, et al. (2000)

Purpose:

Members of the Technical Advisory Committee responded to the conclusions drawn in the previous NAEP evaluation about the appropriateness of the achievement levels that were recommended. Members of the committee were well-known psychometric experts that represented a range of experiences. In addition to Hambleton, Bob Brennan, William Brown, Bill Mehrens, Bob Forsyth, Mark Reckase, Barbara Dodd, Jeff Nellhaus, Doug Rindone, Wim van der Linden, and Rebecca Zwick served as members. Their primary concern was with the NAS's characterization of the standard setting process as "fundamentally flawed" without evidence to support that conclusion. Because this phrase has been cited often in discussions and commentary on the NAEP achievement levels, there are concerns about how this summary conclusion would be interpreted by readers. Another concern expressed by the authors was the absence of a thorough review of the large body of existing literature available on standard setting as it related to their conclusions. The NAS offered an alternative standard setting method in their evaluation study that had not been subjected to peer review or tested in an operational setting. It would be speculative to conclude that the proposed method would have resulted in similar or different results that would have been any more palatable to the evaluators.

Findings:

As a reaction piece, there may not be findings as in a research article. However, it is important to note that anyone who advises on the process will have a clear interest in any conclusions that are drawn from evidence collected in this evaluation. The phrase that the NAS chose to use in their materials to describe the standard setting process would likely elicit a response from any constituency. Any conclusions will require a thorough review of literature and strong evidence to support them.

Implications for the evaluation:

Relevance to the current evaluation:

It will be important to evaluate the current standard setting methodology used for NAEP against existing literature in standard setting. Because the methodology is apparently new, we will probably want to evaluate it's appropriateness against professional literature that describes desirable characteristics of any standard setting workshops (e.g., Hambleton, Berk). This will be important because it is a new methodology that has probably not been submitted for peer review. Another critical component when discussing the standard setting methodologies is to focus on the policy decisions that are made from a standard setting workshop's data. Kane and others have consistently commented on the ultimate policy nature of any cut score or achievement level. This is likely an important educational opportunity for most stakeholders involved.

Title: The Nation's Report Card: Evolution and Perspectives

Authors: Jones and Olkin (2004)

Purpose: This book discusses the history and evolution of NAEP from conception to current issues from the perspective of many individuals who were involved in many of the developmental phases of the project. This 19-chapter text includes discussions on the pioneering work of innovators like Tukey and Tyler, the first administration of NAEP, the development of NAGB, and recent technical advances that have impacted the design, administration, and reporting of NAEP. While much of the book provides anecdotal accounts of the history of NAEP, it offers a unique perspective on the technical and political issues faced by those involved in the process.

Findings:

This text discussed issues relevant to NAEP from a variety of perspectives. Rather than list the numerous conclusions that were drawn by each of the contributors, issues raised that are relevant to the current evaluation design are highlighted below.

Relevance to the current evaluation: Several issues discussed in this book are relevant to the current design for the evaluation of NAEP.

1. Although NAEP was designed to assess large populations, there is a current need assess on a smaller scale, specifically districts and schools. To be useful on this level, multiple policy makers have cited the need for NAEP to produce faster and easily comprehendible results. In addition, numerous authors within the text have cited the lag time in reporting as the top issue that needs to be addressed. The audit study will examine the process of NAEP scoring and reporting.
2. Several authors raise the issues of using NAEP as a monitor for other Federal programs. This was a major consideration in the creation of the current evaluation plan; the use of NAEP specifically to aid NCLB efforts. Overall, each of the studies contributes a unique piece of information for this larger purpose.
3. The inclusion of special needs students has been a constant issue for NCES and NAGB. The biggest challenge in this area is how to incorporate the results for these students with other results. The audit study will incorporate a sampling component which will allow for examination of inclusion policies regarding special populations.
4. One author noted that because of the increased importance placed on NAEP by NCLB, NAGB may not be able to maintain much of the decision making control it once had. As a result, NAGB may have to allocate some of the decision making responsibilities to contractors to maintain the success of the program. The structure of policy and decision making within the NAEP system will be documented as part of the practices evaluated in the Audit study.
5. It was suggested that NAEP could contribute to the NCLB program by providing a measure of progress towards the NCLB goals if the focus of NAEP content was matched to classroom content on the national level. As noted above, the studies in this design each contribute to this overall goal of using NAEP for NCLB

purposes. In regard to this concern, the alignment study will examine how NAEP matches to state content/test frameworks.

6. NAEP results are reported by achievement levels rather than scale scores which are designed to show what students SHOULD be able to do. However, motivation becomes an issue when student participation is not required. One issue not addressed in the current evaluation design is that of student motivation and the effect the lack of motivation may have on performance.
7. NAEP has available a data tool that allows users to search NAEP data and explore results. It is important to determine how this system working and if it allow for easy dissemination of the NAEP results to various stakeholders. One component of the Utility study (study 5 – currently optional pending additional funding) is determining how useful and accessible this information is to interested stakeholders.

Title: Management and technical review of the National Assessment of Educational Progress (NAEP)

Organization: KPMG Peat Marwick LLP (1996)

Author: Huff

Purpose:

It appears that NCES commissioned this external evaluation. Although the title suggests a two-pronged focus, the experience of the reviewing organization weighted the review more heavily to the management and business-related interests of the program. The four areas that the review encompassed included different funding mechanisms for contractors involved with NAEP, costs and monitoring, decision-making processes, and appropriateness of NAEP statistical methods. Although the review was conducted almost a decade ago, many of the issues may still be relevant as it relates to the relationship between NCES and NAGB. The review also mentioned that for specific activities there was the appearance of different philosophies between these agencies that impact the efficiency of reporting NAEP activities.

Findings:

The reviewer produced recommendations from each of the four topical areas. They recommended using cooperative agreements with greater specification of requirements of the agreement and limitations of use to the technical data. There was also the perception that grants to conduct NAEP activities were only available to a few organizations (i.e. ETS and WESTAT). Additional recommendations in this area offered procedural suggestions about how to increase competition through different strategies.

In the second topic area, the reviewer concluded that accounting and record-keeping models used by the contractors were reasonable and met the needs of NCES. It appears, though, that some modifications were made to these record-keeping activities during the review based on earlier requests by NCES.

Within the decision-making section, there were three broad issues that will impact subsequent evaluations. First, the reviewer noted a lack of mission clarity in terms of defining the scope of NAEP and appropriate uses for the scores. Second, there was an observed confusion in management structure as NAGB's unique policy role with respect to NCES's responsibility to conduct NAEP. This continues to be an issue as an examination of these roles was requested by NCES in early discussions for the current evaluation. Finally, the consensus approach for decision-making was one that the reviewer believed appeared to contribute to uncertainty and cost.

As with the decision-making topic, within the statistical methodology section, there was a recommendation for clarifying the scope and purpose of NAEP. A secondary portion of that recommendation included modifying the background questionnaire that is part of the administration. There was also a strong recommendation to shift the testing program from classical testing methods to computer adaptive testing with a belief that changing the mode will impact the motivation of students taking the test.

Relevance to the current evaluation:

The information on the organizational structure and dynamic was very helpful and will be useful when discussing similar questions for the audit phase of the study. Because the reviewer(s) in this evaluation did not appear to have psychometric backgrounds, the conclusions drawn about methodologies and practices did not rely on sufficient, supporting technical evidence. The recommendation to shift the program to a computer adaptive test did not contain an adequate discussion of the methodologies, implications, costs, or a rationale for how changing the mode of administration would increase the motivation for students to participate. There are a number of other considerations that would need to be discussed for an agency to consider this recommendation. The general conclusion for the statistical methodologies that the review examined was that they were adequate. Again, this conclusion did not appear to be based on a strong understanding of the methodologies or the literature on which they were based. An example of this is that the reviewer suggested that DIF analyses could be eliminated for a potential cost savings because ETS indicated that there were few items that were removed from the analysis. This recommendation is inconsistent with professional standards for educational testing. There were also comments related to the NAEP reports that suggest that the question of utility still remains and is likely related to the defined purpose of the program. Articulating the purpose and appropriate uses of NAEP scores will be the most critical step in determining which supporting studies should be conducted. Although the KMPG report would not be considered a full psychometric audit, it does ask some of the broad organizational questions that are included. Additional information is needed about the HumRRO Quality Assurance project to further examine the external review efforts that NCES and NAGB have conducted.

Title: Using the National Assessment of Educational Progress to Confirm State Test Results

Organization: National Assessment Governing Board (2002)

Authors:
Ad Hoc Committee: Nettles, Domenech, Haertel, Kopp, Paulson, Ravitch, Ward, Whirry, & Wolf
Planning Work Group: Reckase, Behuniak, Francis, Holland, Jenkins, LeTendre, M. Shelton, & Yen
Governing Board Staff, Fields

Purpose: The purpose of the report was to provide recommendations and guidelines for using data obtained from NAEP to confirm state results in grades 4 and 8 in reading and mathematics.

Findings: The authors recommended the following seven principles in governing the use of NAEP data to confirm state testing results in reading and mathematics revolving around test development, sampling, data collection and reporting. First, NAEP is well suited to provide confirmatory evidence of general trends in state achievement results (in fact, several states currently use NAEP informally to confirm testing progress). Second, the confirmation should not be viewed as a strict validation of the state's test results due to many limiting factors (some of which were incorporated into the following governing principles). Third, the greater disparity between the state tests and NAEP merit more cautious interpretations about the confirmatory evidence. Fourth, to provide a stable environment for measuring change, the NAEP testing framework should be held consistent for at least ten years, unless an immediate revision is warranted. Moreover, the National Assessment should continue to measure students' knowledge and skills rather than pedagogical methods. Fifth, subgroups should be defined in a similar manner as the state data, sampling procedures should be chosen so that major subgroups produce reliable results, and the exclusion rates on NAEP should be comparable to the state tests. Sixth, the reported NAEP results should be useful to the states and should contain at least the following information: means, percent at or above each NAEP achievement level, percentiles, quintiles, achievement distribution charts, and achievement distribution gap charts. Seventh, the full range of state standards should be considered when evaluating the NAEP achievement levels as confirmatory evidence.

Relevance to the current evaluation:

This study will be a useful reference for the NAEP-State alignment study.

Title: A Content Comparison of the NAEP and PIRLS Fourth-Grade Reading Assessments

Organization: National Center for Educational Statistics (2003a)

Authors: Binkley and Kelly

Purpose: In 2003, NCES will be releasing results for both the 2001 PIRLS fourth-grade assessment and the 2002 NAEP fourth grade reading assessment. In anticipation of questions about comparisons among these assessments, NCES convened an expert panel to compare the content of the PIRLS and NAEP assessments and determine if they are measuring the same construct. This involved a close examination of how PIRLS and NAEP define reading, the texts used as the basis for the assessments, and the reading processes required of students in each.

Findings: The comparison of the NAEP and PIRLS fourth-grade reading assessments suggests that there is a great deal of overlap in what the two assessments are measuring. While they do seem to defining and measuring the same kind of reading, the PIRLS assessment was considered to be easier than NAEP, with more text-based tasks and shorter, less complex reading passages.

The similarities of the assessments are as follows: both define reading similarly (as a constructive process); both assess reading for a literary experience and reading to be informed; both call for students to develop interpretations, make connections across text, and evaluate aspects of what they have read; both use authentic texts as the basis for the reading assessment; and both use multiple-choice and constructed response questions with similar distributions of these types of questions (about half of the items in both assessments are constructed-response items).

The differences of the assessments are as follows: PIRLS calls for more text-based interpretation, while NAEP places more emphasis on having students take what they have read and connect to other readings or knowledge and to critically evaluate what they have read. Close to 20% (18%) of the items in PIRLS require students to locate information in the text that is virtually an identical match to what is in the stem of the item. NAEP does not have any items requiring a verbatim match. PIRLS reading passages are, on average, about half the length of the NAEP reading passages. PIRLS passages are, on average, about 547 words, while NAEP passages are, on average, about 1000 words. Results of readability analyses suggest that the PIRLS reading passages are easier than the NAEP passages (one to two grade levels lower, on average).

Relevance to the current evaluation:

This study is relevant to both the achievement levels studies and the alignment studies. For the achievement levels studies, it will help us describe the similarity in content across NAEP and PIRLS reading tests in fourth grade. For the alignment study, it provides a methodology of comparing test content across two related tests, designed for different purposes and populations.

Title: NAEP Quality Assurance Checks of the 2002 Reading Assessment Results for Delaware

Organization: National Center for Educational Statistics (2003b)

Author: Hoffman, Becker, and Wise

Purpose:

Based on standard analyses of the reading assessment data, NCES commissioned HumRRO to examine specific technical questions to respond to observed outlier characteristics of Delaware's exclusion rates and reading gains for an ethnic subgroup. The questions focused on sampling, case weighting, administration protocol, scoring, scaling and equating, coding demographic information, and test security. HumRRO conducted additional analyses to examine the observed gain scores by recalculating earlier scores using an alternate sample of students that had similar characteristics.

Findings:

For these primary questions, HumRRO's examination of NCES's processes and procedures did not find evidence to suggest that these issues contributed to the observed results. The observed results before the recalculation actually produced larger gains than the original data. However, because the additional analyses were based on small sample sizes, the standard errors were large producing wide confidence intervals. The report also suggested that there were some information about the sampling conditions that were not made available until many of the analyses had already begun. The conclusion was that the results for this subpopulation should be flagged and interpreted with caution. Implications for evaluation:

Relevance to the current evaluation:

This study provides evidence of external review and will be part of the evidence that will inform the audit. Because the contractor noted they did not receive full detail of a sampling question, this may lead to additional questions about the ability to access this information to verify findings. It will be important to discover whether these studies are routinely conducted as an ongoing external monitoring process or commissioned because of concerns with the data in this state. If these types of studies are routinely conducted under the HumRRO contract, it may also provide information on the evaluation of the sampling and weighting methodologies. Another implication is for the ability of NAEP results to confirm state assessment results for NCLB. Given the sample sizes for subgroups in the current design, there may be questions about the utility of scores with wide confidence intervals.

Title: The Measurement of Instructional Background Indicators: Cognitive Laboratory Investigations of Fourth and Eighth Grade Students and Teachers to Questionnaire Items

Organization: National Center for Educational Statistics (2002)

Authors: Levine, Huberman, and Buckner

Purpose: In this study, cognitive interviews (primarily think-alouds) were conducted with 4th and 8th grade students and teachers to determine whether they were correctly interpreting and answering the background questions asking about instructional practices and teacher professional development included on the NAEP reading, math, and science assessments.
Analyses were also conducted in which student and teacher responses to the same question were compared to determine the degree of discrepancy.

Findings: The findings of the study were disappointing with regard to the validity of information obtained from these background questions. High levels of discrepancy between teacher and student responses were found for the majority of questions, and interviewing revealed several sources of confusion. These included:

- Difficulties in estimating how often an activity (such as doing math problems from textbooks or talking about math solutions with the class) takes place. This was especially difficult for students.
- Use of inappropriate response options in the questions.
- Comprehension problems such as whether the phrase "students in your class" referred to the *typical student* or to *any student*.
- Inability to generalize from examples (i.e., taking the example as the only instance that should be counted).
- Loss of context (responding in terms of school experiences in general when the question asks only about math teaching)

Relevance to the current evaluation:

If any studies in the final evaluation design include information from these questions, consideration should be given to these findings and other data sources should be explored. If that is not feasible responses to most background questionnaire items should be treated with extreme caution, and those doing the study should refer to this study to determine the accuracy of responses to the items they are considering. I should point out that information in this study that was obtained from interviews was based on fairly small numbers of teachers and students (12 teachers and 66 students). However, the study also looked at discrepancies for cases in which the same question was asked of teachers and their students and analyses of this data was based on 557 classrooms consisting of approximately 6500 students.

Title: Raising Achievement and Reducing Gaps: Reporting Progress Toward Goals for Academic Achievement in Mathematics

Organization: National Educational Goals Panel (2002)

Author: Barton

Purpose:

The purpose of the study was to examine progress in mathematics educational achievement towards the goals set forth by the Education Summit in Charlottesville, Virginia, in September of 1989. Goal 3 stated that "…all students will leave grades 4, 8, and 12, having demonstrated competency in challenging subject matter…" with the objective that "the academic performance of all students at the elementary and secondary levels will increase significantly in every quartile, and the distribution of minority students in each quartile will more closely reflect the student population as a whole." Using NAEP math scores in grades 4 and 8 from the baseline year to 2000, the study reports state-by-state results regarding changes in: average scores, scores for students in the bottom and top quartiles, percentage of students scoring at or above the proficient level, gap between the scores of students in the top and bottom quartiles, gap between minority and majority student scores, and gap between the scores of students who are and are not eligible for free or reduced lunch.

Findings:

Overall, most states and the nation as a whole showed statistically significant improvement in average scores, scores for students in the bottom and top quartiles, and the percentage of students scoring at or above the proficient level. States were less successful, however, in reducing the gap in the scores between the top and bottom quartiles. Furthermore, there was no significant progress in reducing the gap between minority and majority student scores, and the gap between the scores of students who are and are not eligible for free or reduced lunch.

Relevance to the current evaluation:

This study could be a useful reference in our evaluation of the standard setting on the 2005 NAEP Grade 12 Math Assessment. Given that it emphasizes the importance of valid comparisons across time, it may also be a useful reference for our population invariance studies.

Title: Grading the Nation's Report Card: Evaluating NAEP and Transforming the Assessment of Educational Progress

Organization: National Research Council (1999)

Author: Jones, Mitchell, and Pellegrino

Purpose: Evaluated NAEP's mission and measurement objectives; sampling design and analysis strategies; and the reporting and utility of NAEP's results.

Findings:

1. NAEP serves as a limited by important monitor of academic progress of students in US schools. In order to better meet it's goals, a redesign should be implemented that uses additional assessment resources to monitor academic progress.
2. NAEP's design is overly complex. Recommendations are made to streamline NAEPs design through integrated assessment systems.
3. Strategies are needed to standardize and enhance the participation in NAEP by students with disabilities and English language learners.
4. NAEP assessments do not capitalize on contemporary research, theory and practice in ways that support interpretations of student knowledge and understanding.
5. NAEP achievement levels are fundamentally flawed due to the ways they are set and the lack of validity evidence to support their use.

Relevance to the current evaluation:

1. Several reports are referred to (NAEP Technical Reports, for example) that contain information about the psychometric properties of the results. These reports should have relevant information for the audits.
2. Several studies are referenced about the presentation and interpretability of NAEP reports. NAE has articulated criteria for NAEP reports (accuracy of results, likelihood results would be interpreted correctly by the intended audiences, extend to which the results are accessible and adequately disseminated, and timeliness with which results are made available); these criteria could serve as components for the study examining the interpretation of NAEP reports.
3. Some, very limited, information is provided about the current sampling plan (p. 66-67; 68).
4. Several areas were pointed out that need attention: a) sampling designs (problem with administration burden for small and low density states and large school districts and the analytical complexity that limits the number of items per examinee), b) complexity of the current analytical methods (including the use of imputed plausible values and the length of time it takes to get results due to the computational complexities); c) calculation of the standard error (the estimation procedure results in an underestimation of the standard errors), d) unreasonably short time frame for test development.

5. Attention to inclusion of all students highlights the need for a study of population invariance of the results.

Title: A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA)

Author: Nohara (2001)

Purpose: This study compares the eighth-grade science and mathematics portions of NAEP 2000 with TIMSS-R (the repeat of the Third International Mathematics and Science Study) and the scientific literacy and mathematics literacy portions of PISA (the OECD's Programme for International Student Assessment) and is intended to help those interested in learning more about the assessments, including their purposes, their similarities and differences, and the relative emphasis each one places on the various content areas and types of knowledge. It is based on the work of expert panels in science and math education and testing who analyzed each assessment item in various categories. It is not intended to facilitate the translation of performance on one of the three into a projected performance on one of the others, nor is it intended as an evaluation of the quality of any of the assessments. However, this report should help those wishing to understand the differences between the three assessments and how they might influence performance.

Findings: Expert panels in mathematics and science education examined items on each of the three assessments in terms of content, response type, context, requirements for multi-step reasoning, and other characteristics. For all of the characteristics except content, the panels used sets of descriptors developed specifically for this comparison. In the area of curriculum content, panel members compared the three assessments to the NAEP "Fields of Science" and mathematics "Content Strands." The assessments were thus compared using a set of common criteria, which, in almost all cases, were different from the criteria used to develop each assessment. Because each assessment was developed based on a different underlying philosophy and set of frameworks, and while sharing many common characteristics, the assessments each have different emphases on content and item type. In both science and mathematics, there are significant differences between the assessments in most areas examined, many of which can be traced to differences in the purpose of each assessment. Both NAEP and TIMSS-R seek to assess students' mastery of basic knowledge, concepts, and subject-specific thinking skills tied to extensive frameworks of curriculum topics. As a result, both assessments have large numbers of items covering a broad range of topics, with items generally focused on a single, identifiable piece of knowledge, concept, or skill. Some items draw on a combination of topic areas or are more focused on students' scientific or mathematical thinking abilities than on content topic, but these items were in the minority. In contrast, the purpose of PISA is to assess students' abilities to handle everyday situations that require scientific and mathematical skills. As a result, PISA items fit less well on frameworks of curriculum topics and are more often set in real-world contexts.

Relevance to the current evaluation:

This study is relevant to both the achievement levels studies and the alignment studies. For the achievement levels studies, it will help us describe the similarity in content across NAEP, PISA, and TIMSS tests. For the alignment study, it provides a methodology of comparing test content across two related tests, designed for different purposes and populations.

Title: Trends in Large-Scale Testing Outside of the United States

Author: Phelps (2000)

Purpose:

This article focuses on concerns about testing trends internationally. In response to Congressionally funded studies that suggested there was a decrease in large-scale testing in other countries, the author provides information on changes that have occurred in testing internationally over the past 25 years. In his examination of 31 countries, he lists tests that have been added or dropped from testing programs during this time period. In addition, he describes the purpose of these tests, most of which have been exit or entrance examinations. Another area of testing that has begun to increase more systematically is the international tests such as the Third International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA). These testing programs are generally used for comparing student performance across countries.

Findings:

The key finding from these analyses was that the testing trend with respect to these types of tests has not declined rather it has increased. An interesting comment that the author does not make, though, is the use of national tests in these other countries for purposes similar to NAEP. It appears that Canada added a similar program in 1993, the Czech Republic was developing a system in 1997, England and Wales added a program in 1991, France added a system in 1980, Japan introduced a program in the 1980s, the Netherlands also added a system in the 1980s, Spain added a system in 1996, and Sweden added a system in 1989. In each of these instances, there are no "stakes" associated with these exams. The entrance and exit tests described have greater similarities to state graduation demonstration tests or college entrance exams (i.e. SAT, ACT) in the United States. It is possible that these tests are used in other countries, but because the analyses focused on changes (additions or deletions) in testing programs over the past 25 years, these types of tests may not have been identified.

Relevance to the current evaluation:

Trends of testing programs added internationally appear to support information that is being used directly for educational decisions. Thus, this information may guide our interpretations of the range of purposes for NAEP scores. It also suggests that studies examining the utility of information will be important because the tests similar to NAEP used in other countries are likely informing progress on a nationally mandated curriculum. It is possible that alignment studies may be needed to determine the extent to which NAEP scores inform state content standards or state assessment scores inform the NAEP frameworks. Each of these supporting studies would address different questions. Some of these studies may already be available. However, it leads to the broader question of whether NAEP can be used to validate state assessment information.

Title: NAEP Background Questions: What Can We Learn from NAEP About the Effects of Schools and Teachers on Student Achievement?

Author: Podgursky (2002)

Purpose: This is a discussion paper prepared for NAGB that questions the utility of most NAEP reports specifically with respect to inferences based on the background questions. The author suggests that due to significant measurement error in responses to these questions (as reflected in comparisons between Census and NAEP data, and other such sources) that these data do not provide an adequate basis for the kinds of near-statistically significant inferences made by authors of many NAEP reports (i.e., one small step short of identifying causal relationships). The author's main concern involves the extent to which NAEP reports are relied on for answers for ways to raise student achievement (seemingly without regard for whether that is an intended or unintended use of the data: since the questions are there, that is what people are using the data for).

Findings: Podgursky identifies several limitations to the NAEP background questionnaires, which in sum are more generally reflective of the limitations of gathering information on SES and other similar background variables in most any context in the social sciences. Sources of problems are measurement errors in teacher and student responses and missing data, the lack of longitudinal data, and confounds in teacher quality and classroom behavior variables.

One proposition clearly made in this paper is that many NAEP reports are more akin to policy documents written to advance a particular point-of-view with respect to the data (for example, that NAEP authors sometimes downplay evidence inconsistent to the perspective they want to advance to highlight more favorable evidence, as well as selectively drive the choice of variables for this purpose). The solution advanced by Podgursky is to publish most background data in table form without comment.

An issue also raised here is the use of imputed values and not direct measures of student scores. He notes that other authors have similarly expressed difficulties with plausible values, but offer no unique criticisms or suggestions for alternatives.

Relevance to the current evaluation: Moderate to low. This report will be useful if the utility of NAEP reports study is funded. NAEP collects a great deal of information from students and teachers alike, and since time and money are spent to develop these questions and to report on the results, independent evaluation of the results is warranted. This article raises several issues that are important to this process with respect to measurement of variables and how the way a question is asked frames the nature of the response. At the same time, however, the comparisons Podgursky makes between research on NAEP and experimental studies in medicine fall a bit short in that while the scientific method is what it is, and is the ideal to strive for in all research, determining the effects of a pill on a cold is still a different enterprise than determining the impact (for example) of the impact of a teacher on learning gains of students.

Title: Contributions of Background Questions to Improving the Precision of NAEP Results

Author: Reckase (2002)

Purpose:

This paper provides a general overview of the conditioning process and the production of plausible values for estimating distributions of student proficiency.

Findings:

The importance of the background variables varies based on the content representation of the booklets. If the booklets are content balanced, the background variables have more influence on the proficiency estimation. Recommendations are made regarding how to determine which background variables should be used. While there is a desire to maintain parsimony given the large number of possible variables, the exclusion of variables that seem unrelated to the proficiency estimates results in a more biased estimate, if different conditioning variables were used from year to year.

Relevance to the current evaluation:

This document is valuable in that it provides a good overview of the estimation of the proficiency distribution and provides some practical guidelines in choosing the conditioning variables. Thus, it is particularly relevant to the population invariance studies. If there is a need to evaluate the process of computing the plausible values, this is a paper that provides guidance in doing so.

Title: The NAEP 1998 Technical Report

Organization: National Center for Educational Statistics (2001)

Author: Allen, Donoghue, and Schoeps

Purpose:

This report carefully details the entire process for the NAEP 1998 assessment beginning with the development of the test framework through administration and specific analyses conducted for each subject area. The authors include in the purpose statement, that the report demonstrates adherence to professional standards for testing practices (e.g., *Standards for Educational and Psychological Testing,* AERA, APA, NCME, 1999). This document is a nice reference for anyone seeking specific information about any stage of the NAEP process. In addition, this document contains much detail that does not appear to be available to the public elsewhere (e.g., description of the sampling weighting process).

Relevance to the current evaluation:

The technical documentation contained within this report will be highly relevant to the current evaluation and will inform some if not all of the studies on the background of NAEP methodologies. Suring the review process of this and other documents, it became apparent that much of the work described in this report is based on policy that has since changed due to innovations in testing practices or changes in Federal legislation, for example. This document will be used as a reference for past policy during the evaluation (i.e., when addressing specific aspects of the NAEP assessment program in the audit) and may inform specific questions regarding change in policy and procedure.

Title: Overseeing the Nation's Report Card [subtitle: The Creation and Evolution of the National Assessment Governing Board (NAGB)

Author: Vinovskis (1998)

Purpose: To provide an historical summary of NAEP and NAGB from the inception of NAEP through the mid 1990s. The study provides some focus on controversies related to setting achievement levels (performance standards).

Findings: There are no specific findings of this study other than the historical representation of how NAEP and NAGB came into being, some discussion of controversies, and aspects of policy issues that reflected on the development of selected aspects of NAEP through the mid 90s.

Relevance to the current evaluation: This document provides some insights about why NAEP, NAGB, and NCES relate to each other the way they do (assuming some constancy in these relationships over the past 10 years or so). The purpose and responsibilities of the governing board are described and will be useful in the audit study as a precursor to the site visit to NAGB. The description of the achievement levels process will be useful in preparation for the validity of achievement levels study. This section provides a summative history of the controversy and issues surrounding setting the achievement levels.

Reference List

Abedi, J., Lord, C., Hofstetter, C. (2001, September). *Impact of Selected Background Variables on Students' NAEP Math Performance.* Paper prepared for the U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, Washington, DC.

Abedi, J., Lord, C., Kim, C., Miyoshi, J. (2001, September). *The Effects of Accommodations on the Assessment of LEP Students in NAEP*. Paper prepared for the U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, Washington, DC.

American Institutes for Research. (2004a, May). *Evaluation of Bias Correction Methods for "Worst-case" Selective Non-participation in NAEP*. Palo Alto, CA: McLaughlin, D., Gallagher, L., Stancavage, F.

American Institutes for Research. (2004b, May). *Federal Sample Sizes for Confirmation of State Tests in the No Child Left Behind Act*. Palo Alto, CA: Mosquin, P., Chromy, J.

American Institutes for Research. (2002a, October). *An Agenda for NAEP Validity Research*. Palo Alto, CA: Stancavage, F.B., Beaton, A.E., Behuniak, P., Bock, R.D., Bohrnstedt, G.W., Champagne, A. et al.

American Institutes for Research. (2002b, September). *National Longitudinal School-Level State Assessment Score Database. Analyses of 2000/2001 School-Year Scores.* (Contract: ED-01-CO-0026 0002). Palo Alto, CA: McLaughlin, D., Bandeira de Mello, V., Cole, S., Blankenship, C., Hikawa, H., Farr, K., et al.

American Institutes for Research. (2002c, July). *The Validity of Oral Accommodation in Testing*. Palo Alto, CA: Weston, T. J.

American Institutes for Research. (1999, March). *An Investigation of Why Students Do Not Respond to Questions*. Palo Alto, CA: Jakwerth, P.M., Stancavage, F.B., Reed, E.D.

American Institutes for Research. (1998a, September). *Reporting the Results of the National Assessment of Educational Progress.* Greensboro, NC: Jaeger, R.M.

American Institutes for Research. (1998b, August). *The Effects of Finite Sampling on State Assessment Sample Requirements*. Palo Alto, CA: Chromy, J.R.

American Institutes for Research. (1997a, December). *A Study of Equating in NAEP*. Palo Alto, CA: Hedges, L. V., Vevea, J. L.

American Institutes for Research. (1997b, May). *Optimizing State NAEP: Issues and Possible Improvements.* Palo Alto, CA: Mullis, I.V.S.

Barton, P. E. (2002, September). *Perspectives on Background Questions in The National Assessment of Educational Progress.* Paper prepared for a conference sponsored by the National Assessment Governing Board, Washington, DC.

Bourque, M.L. & Byrd, S. (Eds.). *Student Performance Standards on the National Assessment of Educational Progress: Affirmations and Improvements.* Washington, DC: National Assessment Governing Board, 2000.

Bourque, M.L. (Ed.). *Proceedings of Achievement Levels Workshop*. Washington, DC: National Assessment Governing Board, 1998.

Educational Testing Service. (2004). *Including Special-Needs Students in the NAEP 1998 Reading Assessment: Part II, Results for Students with Disabilities and Limited-English-Proficient Students.* (ETS-NAEP 04-R01) Princeton, NJ: Lutkus, A.D., Mazzeo, J., Zhang, J. & Jerry, L.

Educational Testing Service. (2003a, November). *Growth in School Revisited. Achievement Gains from the Fourth to the Eighth Grade.* Princeton, NJ: Coley, R.J.

Educational Testing Service. (2003b, November). *Analysis of NAEP Combined National and State Samples.* Princeton, NJ: Qian, J., Kaplan, B., Weng, V.

Educational Testing Service. (2003c, April). *The Impact of Item Treatments on NAEP Reporting Scale Scores.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Educational Testing Service. (2003d, April). *Statistical Power Analysis and Empirical Results for NAEP Combined National and State Samples.* Princeton, NJ: Qian, J.

Educational Testing Service. (2001, February). *Differences in Gender Gap: Comparisons Across Racial/Ethnic Groups in Education and Work.* Princeton, NJ: Coley, R.J.

Grissmer, D. (2002, September) *Improving NAEP for Research and Policymaking.* Paper prepared for a conference sponsored by the National Assessment Governing Board, Washington, DC.

Hambleton, R.K., Brennan, R.L., Brown, W. J., Dodd, B., Forsyth, R.A., Mehrens, W.A. et al. (Summer 2000). A Response to "Setting Reasonable and Useful Performance Standards" in the National Academy of Sciences' Grading the Nation's Report Card. *Educational Measurement: Issues and Practice*, 5-15.

Jones, L.V., & Olkin, I. (Eds.). (2004). *The Nation's Report Card Evolution And Perspectives.* Bloomington, IN: Phi Delta Kappa Educational Foundation.

KPMG Peat Marwick LLP. (1996, June). *Management and Technical Review of the National Assessment of Educational Progress (NAEP)* (study). Washington, DC: Huff, L.D.

National Assessment Governing Board. (2002). *Using the National Assessment of Educational Progress to Confirm State Test Results* (A Report of The Ad Hoc Committee on Confirming Test Results). Washington, DC: Government Printing Office**.**

National Center for Education Statistics. (2003a). *A Content Comparison of the NAEP and PIRLS Fourth-Grade Reading Assessments. Working Paper Series.* (NCES-WP-2003-10). Washington, DC: Binkley, M., & Kelly, D.L.

National Center for Education Statistics. (2003b). *NAEP Quality Assurance Checks Of The 2002 Reading Assessment Results For Delaware.* (WP-2003-10). Washington, DC

National Center for Education Statistics. (2002) *The Measurement of Instructional Background Indicators: Cognitive Laboratory Investigations of the Responses of Fourth and Eighth Grade Students and Teachers to Questionnaire Items.* (NCES 2002-06). Washington DC: Levine, R., Huberman, M., Buckner, K. Goldstein, A.

National Education Goals Panel. (2002, January). *Raising Achievement and Reducing Gaps: Reporting Progress Toward Goals for Academic Achievement in Mathematics*. Washington, DC: Barton, P.E.

National Research Council. (1999, January). *Grading the Nation's Report Card: Evaluating Naep and Transforming the Assessment of Educational Progress.* Washington, DC: Jones, L.R., Mitchell, K.J., & Pellegrino, J.W.

Nohara, D. (2001, June). *A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMISS-R), and the Programme for International Student Assessment (PISA).* Paper prepared for the U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, Washington, DC.

Phelps, R.P. (Spring 2000). Trends in Large-Scale Testing Outside the United States. *Educational Measurement: Issues and Practice*, 11-21.

Podgursky, M. (2002, September) *NAEP Background Questions: What Can We Learn from NAEP About the Effect of Schools and Teachers on Student Achievement?*

Discussion paper prepared for the National Assessment Governing Board, Washington, DC.

Reckase, M. D. (2002, September). *Contributions of Background Questions to Improving the Precision of NAEP Results.* Paper prepared under contract to the National Assessment Governing Board, Washington, DC.

U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics. *The NAEP 1998 Technical Report*, NCES 2001-509, by Allen, N.L., Donoghue, J.R., & Schoeps, T.L. (2001). Washington, DC: National Center for Education Statistics.

Vinovskis, M.A. (1998). *Overseeing the Nation's Report Card. The Creation and Evolution of the National Assessment Governing Board (NAGB).* University of Michigan, Department of History, Institute for Social Research, School of Public Policy.