

# **Independent Audit Panel Review of North Carolina Testing and Accountability Issues**

presented to

North Carolina  
State Board of Education

December 2001

# Table of Contents

- Executive Summary 1
- Introduction 13
- Background 14
- Findings and Recommendations 16
  - Findings on the end-of-grade (EOG) assessment program 16
  - Recommendations on the end-of- grade (EOG) assessment program 19
  - Findings on general testing and accountability issues 20
    - Resources 20
    - Time and effort 22
    - Curriculum and assessment 23
    - Standard-setting 24
  - Recommendations on general testing and accountability issues 25
  - Findings on general decision-making issues 27
  - Recommendations on general decision-making issues 28
  - Findings on key groups 30
  - Recommendations on key groups 32
  - Findings on technical quality and clarity of information 33
  - Recommendations on technical quality and clarity of information 33
  - Findings on oversight of testing and accountability 33
  - Recommendations on oversight of testing and accountability 34
- Align360's 35
  - Additional recommendations on general testing and accountability issues 35
  - Additional recommendations on general decision-making issues 36
  - Additional recommendations on key groups 36
  - Additional recommendations on technical quality and clarity of information 37
- Conclusions 38
- Attachment A 40
- Attachment B 41
- Attachment C 44

# Executive Summary

## Conclusions

---

The Audit Panel is impressed with the North Carolina assessment and accountability program and its role to increase the quality of educational programs and student achievement in the state. Apparently, in North Carolina and other states, the support for a comprehensive approach to education improvement including teacher salary increases and school facilities funding would not have occurred without a strong testing and accountability program that showed improvement in student achievement results.

North Carolina policy-makers, especially the State Board of Education, the General Assembly, and the Governor, should be recognized for their support for the variety of approaches, including testing and accountability, to raise the achievement of all students.

The current testing framework of end-of-grade tests in grades 3-8 and end-of-course testing for high school is an appropriate structure for testing and accountability. The approach has contributed significantly to North Carolina gains in student achievement as reflected in the National Assessment of Education Progress and other external measures, as well as being a model for other states.

The setting of mathematics cut scores for end-of-grade tests illustrates that important issues are present and must be addressed to assure the integrity and credibility of the testing and accountability program each and every year.

These primary factors contributed to the recent problems with the mathematics test:

- An implementation timetable that was too short. No time was available for a structured review of results to ensure adequate technical accuracy.
- Inadequate resources and staffing. New tests and new purposes for testing were added faster than resources and staff were added to do the work.
- Major changes were made too frequently to content standards. Significant changes in a short period of time spell trouble in test development and standards setting.
- Inadequate communication to, and direct involvement of, the State Board of Education in setting testing and accountability standards.

The Department of Public Instruction testing and accountability staff is dedicated and hard working. The staff has been over extended in meeting the increased demands on testing and accountability through the 1990s by the State Board of Education, the Governor's office, and the Legislature. The staff turnover rate is an unmistakable signal of serious problems with insufficient staffing and resources that have been cited repeatedly.

The assessment program is not funded adequately, and needs greater coordination between the various interested parties. Design issues are present that contributed to the scoring problem in May 2001. Unless the end-of-grade testing and end-of-course testing program is modified as suggested in this report problems of this nature will likely re-occur.

Assessment programs cast in a framework of high-stakes decisions must be psychometrically and legally defensible for they are invariably challenged on both counts. Although the planned North Carolina graduation test was not a direct topic of study, significant concerns were expressed by the panel about the present plans for the test.

The stakes for the graduation test are even higher than using tests for promotion decisions. This test is likely to be challenged in court, and every effort should be made to assure that the state is legally defensible.

No state approach to testing and accountability is perfect. The use and application of state tests for accountability purposes is a relatively recent development. Maintaining the integrity of the testing and accountability program absolutely requires dedicating sufficient resources, appropriate timelines, adequate oversight, and the involvement of key policy-makers.

Adjustments to the testing and accountability program that ensure quality, integrity, fairness, and practical use of tests will help support credibility and continued support for a model tool for improving student achievement.

## End-of-Grade Testing Program Findings ---

Student assessment programs can be designed in many different ways. No single correct design exists. Further, each design can be strengthened (or weakened) by decisions that are made (or not made) as the program is implemented.

As in other states, the North Carolina General Assembly and State Board of Education are interested in using test results in accountability programs as a basis for allocating financial incentives to educators and schools. The EOG program became a high stakes program in which both students and educators had something to risk. In this context, the state must be concerned about the psychometric quality of the tests,

student opportunity to be taught the required content, the quality of the testing process, the maintenance of testing program security, and due process issues. The Audit Panel's view is that DPI is not well positioned currently to be responsive to each of these issues with equal effectiveness.

The Audit Panel was told during the meetings that the new mathematics curriculum introduced in 1999-2000 represented "sweeping," "drastic," "cutting-edge" and "revolutionary" changes. DPI curriculum specialists wanted the new tests to reflect this new curriculum as closely as possible. Such dramatic changes and the associated testing created several psychometric, instructional, and legal problems. Adequate resources, time, and planning are necessary for such transitions.

First, North Carolina law requires that the curriculum changes be based on a five-year renewal cycle. Given the substantial time required to write and field test new items, the assessment program staff is in a constant struggle to keep the tests current. Second, new versions of the tests can out-pace the current instruction in the schools, thus denying students an appropriate opportunity to learn.

In May 1999, DPI contracted with an outside firm to develop approximately 14,000 new test questions for field testing in the May 2000 test forms. This is an enormous number of items for DPI to process in a short time, and the Audit Panel believes it was overly ambitious.

DPI contracted with a local university to prepare test forms. Field testing the large number of items required about 700 forms to be created quickly. The Audit Panel had the impression that the current arrangement with the university was insufficient to meet the demands of the embedded field testing system.

The Audit Panel was unable to isolate why the statistical information coming from the May 2000 field test did not reflect accurately student proficiency and, thereby, incorrectly projected the cut-scores on the new score scale. The Panel was informed that staff had insufficient time to analyze the statistical data regarding students omitting items at the time the problems were discovered.

No independent quality control analysis was performed by DPI to check the work of its UNC-Chapel Hill contractor responsible for the EOG linking analysis based on the May 2000 data. However, the Panel has no information that would suggest errors in the analyses of the May 2000 test results.

DPI has a mechanism that permits retrieval of early test data, but little time is available for staff to revise procedures if problems are identified.

As early as fall 2000, the UNC-Chapel Hill contractor communicated to DPI the possibility (but not the certainty) of a problem with linking of old and new test forms. The North Carolina psychometric Technical Advisory Committee was re-established and met in January 2001 to discuss this and other issues.

No time was available before May 2001 results were released for DPI to evaluate those results to determine if a problem actually existed and, if so, to conduct a new standard setting study. The quandary of the existing program design is attempting to make large changes in teaching and testing within an assessment program that requires both speed and continuity of results.

## **A. Recommendations on End-of-Grade Testing**

1. DPI needs to develop a plan for addressing changes in curriculum expectations to be coordinated with changes in the tests. Either sweeping changes cannot be made or, if made, all persons affected (school personnel, legislators, state government officials, DPI staff, students, parents, and the public) must be informed that comparable test results cannot be provided.
2. DPI should determine the various stages of test development — applicable to any test — and adopt those procedures as policy. A public formulation of policy puts everyone on notice that the state takes the test development and accountability process seriously.
3. The test development process for all tests should have a set of minimum components. These efforts need to be coordinated for subject areas, grade levels, and program components in terms of planning and projecting timelines.
4. All test development and implementation projects should be accompanied with detailed timelines that reveal the consequences of all major decisions related to meeting (or not meeting) deadlines.
5. If the current design of the EOG tests is to remain in place (i.e. districts providing the scoring services), arrangements should be made to obtain a representative data set and review it for accuracy before districts complete the scoring and reporting operation. Appropriate resources should be in place for this review, including appropriate DPI staffing and software support, as well as external consultants and experts (such as those on the Technical Advisory Committee).

# General Testing and Accountability Issues Findings ---

## *Resources*

Policy-makers in North Carolina have mandated uses for student achievement data that have high stakes for students and schools. To support these uses, the testing program must be psychometrically sound and defensible. A program that meets these goals must have adequate funding. Apparently, the financial resources available to DPI for its assessment operations are inadequate for the tasks.

The Audit Panel believes North Carolina should conduct a review that has the support and confidence of legislative leadership to determine how additional resources can be provided to support the statewide assessment program.

Moreover, insufficient personnel are available to the assessment program to complete the assigned tasks. The Department of Public Instruction testing and accountability staff should be commended for their hard work, commitment, and professionalism. The state is fortunate to have a core staff who has done extraordinary work with a high degree of professionalism despite inadequate resources, staff, and support. North Carolina policy-makers need to address the issue of insufficient resources.

## *Time and Effort*

The Audit Panel is concerned that policymakers may lack an understanding of the time and effort needed to produce a testing program that is consistent from year to year and can be used for high stakes decisions. Because of this lack of understanding, schedules are set without sufficient time to guarantee quality tests and to check results carefully before reporting them to the schools and the public. Part of a solution to this problem is to inform policy makers about the realities of development and implementation of high stakes tests.

## *Curriculum and Assessment*

Tension exists between the desires of curriculum leaders and psychometricians. Curriculum leaders, interested in improving the curriculum, want to use large-scale assessment tests to direct change. In contrast, psychometricians worry that any changes in the tests can de-stabilize the statistical relationships.

For an assessment program to have continuity, requirements must be balanced so changes in tests do not result in serious damage to the equating processes and schools can adjust their instruction and students' needs.

The key is a decision-making process that reviews potential changes to test forms considering all issues—equating, content changes, form improvement—and then makes decisions by considering a balance among all of the issues. For statewide assessment programs, considering major curriculum shifts on an eight- to 10-year cycle is appropriate.

From a management point of view, the assessment staff should be in control with curriculum staff providing support. If litigation occurs, the assessment staff will carry the burden of defense. Therefore, the tests must be designed and implemented in a manner that meets psychometric standards.

### *Standard Setting*

DPI should provide careful training for their standard-setting panels and provide as much information as possible to ensure well-informed judgments. The State Board of Education should be involved more directly in the standard setting process and should be ultimately responsible for setting school and student achievement standards that strike the appropriate balance between challenging and rigorous and reasonable and fair.

The pursuit of the question “how good is good enough?” regarding school and student performance standards should be a very public and understandable discussion. The manner and timetable in which the recent math standards were set did not allow for necessary public debate and understanding.

## **B. Recommendations for General Testing and Accountability**

1. DPI should review the organization of its curriculum and assessment operations to create a firm chain of command and assignment of responsibilities. Assessment operations and needs should be paramount considerations.
2. Because resources for administering the North Carolina testing programs have been declining, substantial staff time is devoted to completing tasks with tight deadlines and to crisis management. Greater resources are needed for more emphasis on long-term planning. DPI assessment staff are competent, dedicated and have a laudable “can do” attitude despite increased demands on the testing program and decreased resources.
3. DPI assessment staff need to produce realistic, well-defended plans for resources to conduct successfully the North Carolina testing programs. Funding authorities should provide those resources or reduce the requirements of the testing programs.
4. DPI should initiate plans that provide for quality control operations that are entirely separate from test support contractors. All statistical and scoring work must be independently verified.

5. DPI must ensure that complete documentation of all major aspects of the programs be produced routinely if the program is to be appropriately reviewed and legally defended.
6. DPI could benefit from allowing its TAC to meet more frequently each year, especially in times when changes are anticipated. DPI should consider expanding the membership of the TAC for additional independent advice from people who are not state contractors.
7. Plans should be in place for real-time decision-making during the critical period when test results are being reviewed. Supervisors should be alerted to the fact that the review will take place, be made aware of implications of their decisions, and be given background information necessary to make their decisions.
8. The greater the stakes involved in the testing program, the more important are accuracy and credibility of results. Test scores should not be released for use before passing the approval process.

## Decision-Making Findings ---

Unquestionably, the North Carolina student assessment and school accountability programs have had impact. North Carolina education has benefited from this effort and will continue to do so in the future. At the same time, some confusion and disagreement over objectives and the particular role of the statewide assessments appears to be present.

States vary in the design of their testing programs. The current North Carolina program may or may not meet the needs of the state for the next decade. To review the program and make adjustments as needed is entirely appropriate.

These recommendations seek to help North Carolina establish appropriate test development processes that meet professional standards. Thoughtful efforts to establish the basic purposes of accountability and to identify practices and resources necessary to ensure high quality are fundamental for improvement of the program. Quality test development, standard setting, and implementation of those standards are needed to continue progress in raising student achievement.

### **C. Decision-Making Recommendations**

1. Develop and document processes for test development and standard setting that assure technical quality. Align standards of quality with those of the national pro-fessional testing organizations found in the *Standards for Educational and Psychological Testing* (1999).

2. Review North Carolina statutes and State Board of Education policies to verify that they are coordinated and that they clearly state the purposes to be served by the statewide assessment programs. The design and funding levels of the testing and accountability programs should reflect those purposes and priorities.
3. Review and revise State Board of Education policies that direct test development and standard setting operations. Clarify lines of authority and the role of the advisory committees. The General Assembly should give specific authority and responsibility to the Board to establish the necessary safeguards and processes to assure the technical quality of testing and standard setting and the authority to adopt passing scores.
4. Establish a process for the State Board of Education and public review of student and school performance standards that includes complete and accurate information, provides adequate time, and involves a variety of stakeholders. Ultimately, the decision-making rests on a fully informed State Board of Education to maintain a balance between challenging standards and what is reasonable and fair.
5. Conduct a review of major court cases of student testing programs. Determine whether North Carolina's student assessment programs are designed and documented properly so the state can be defended adequately if litigation should occur.
6. Acknowledge operational limitations of state assessments. Establish a priority for technical quality, reliability, and validity. Balance the limitations of state assessments with practical and sound approaches to support classroom teaching practices and parental need for information. Continue to communicate clearly the purposes and uses of testing and standards to educators and the public.
7. Schedule periodic external reviews by experts in psychometrics and accountability policy for quality test development and standards. Establish a program of research into test development and standard setting procedures to determine how to best align state content expectations, instruction, and assessment and to best promote continued improvement in testing and accountability.
8. Establish reasonable timetables for test development and standards setting (allowing adequate time for appropriate planning and technical procedures) to ensure high levels of technical quality.
9. Establish long- and short-range plans for test development and standard setting. Timetables should be scheduled according to appropriate sequencing and the time necessary to accomplish tasks. Program transitions should be anticipated to ensure stability and continuity.

10. Procedures for major changes in the testing program should be planned carefully at both the state and local levels. Changes in content, testing, and performance standards should be planned so the stability and integrity of the entire program are supported. Timetables should reflect periodic minor changes with less frequent major reviews.
11. The Legislature should clarify the provisions in statute (Section 28.17.h) that limit field-testing of tests. All state or national testing programs must be able to gather representative data through field-tests and statistically evaluate new test questions.

## Findings on Key Groups ---

A lack of a clear definition of responsibilities for the groups involved in planning test development and standards setting is evident. The operation and structure of the Technical Advisory Committee and Compliance Committee should be reviewed and changed. The procedures and relationships are too informal and unstructured. These committees are important to North Carolina if they function well and are linked properly to DPI staff and State Board.

The organizational culture may prevent the collection and distribution full information regarding important issues necessary to the decision-making process. The State Board received a report on this and other related issues on May 29, 2001. This suggests that the State Board of Education was inadequately informed about key technical and operational issues in the mathematics linking. When new important information is available, the structure and the operating “culture” should be in place for that information to get to the Board.

Policy-makers must understand that testing is a technical science as well as an art form. When policy decisions are made that are contrary to the science of testing, the program is weakened and may not be able to support policy goals. In the end, those choices and compromises have to respect the statistical and psychometric science of testing. The State Board must be certain that its policy decisions are always informed by the best technical advice. This means that policy decisions made with “the best information available” must not be viewed by Board and staff as irreversible if better or more complete information becomes available.

Ultimately, test development and standards setting decisions should support the appropriate use of tests and standards for accountability purposes. Striking the balance between the limitations of state assessment and the need to hold schools and students to those standards is a challenge requiring careful consideration and fair implementation.

Cross checks and oversight need improvements to ensure the integrity of the testing program. The reliability and validity of testing are extremely important because high stakes are attached to testing results. Equally important is the accuracy of reporting and analysis of results.

Unfortunately, no regular and systematic approach to external review of testing, standards setting, and other aspects of accountability has been established. The establishment of performance standards for individual students and schools are decisions that specifically require timely involvement of the State Board of Education.

Inadequate time and inappropriate sequencing of activities did not allow for providing appropriate, full, and accurate information necessary for well-informed decision-making. As a result, the State Board of Education was informed in a less than adequate and timely fashion. However, the short timetable for implementation was one of the most important contributing factors to difficulties with inadequate information.

#### **D. Recommendations for Key Groups**

1. Establish realistic approaches for compiling and using results from the testing and accountability program. Establish methods that ensure quality, clarity, timeliness, technical accuracy, and appropriate use of the information.
2. Establish priorities for information considered most necessary for State Board decision-making and provide sufficient time for adequate consideration of the issue(s).
3. Develop regular methods for communicating appropriate information to local educators and state policy-makers including higher education officials, the Governor's Office, and the General Assembly. Educating interested parties with clear, accurate, and timely information is key to sustaining understanding and support of the testing program and standards.
4. Occasional independent reviews of tests and standard setting processes would ensure integrity of the testing, standard setting, and accountability program.
5. Policy-makers should have regular opportunities to see how the testing and standard setting process works. Recent legislation regarding testing and accountability suggests a basic lack of understanding of the testing and accountability program. Efforts need to be made to help legislators understand how the system can be improved and what resources are necessary to make those improvements.
6. Actions by the State Board of Education should be made based on complete and accurate information with full recognition of the consequences of policy decisions. Thorough, straightforward, and accurate information about test development and standards should be provided to the State Board so it can make policy decisions that

reflect an understanding of the operational feasibility and technical and legal limitations of testing and standards.

7. Establish statutes and State Board of Education policies that ensure technical quality, operational viability, and fairness of the testing program. Outline the processes and responsibilities of parties involved in standard setting and test development.
8. Establish responsibilities for advisory groups to the State Superintendent and State Board of Education. Those groups should include local educators and various Department of Public Instruction staff. Also included in the responsibilities of advisory groups should be internal and external review of the processes for making recommendations.
9. Establish process standards for testing, standards setting, and accountability that applies to Department of Public Instruction staff, State Superintendent, and State Board of Education. These standards will provide a base to make decisions that are independent of internal and external political pressures. The standards will help support agency staff to provide professional advice that fully informs policy-makers of the impact of their decisions.

# Report of the Independent Audit Panel

# Introduction

In July 2001, North Carolina State Board of Education (SBE), with the approval of the SBE ad hoc committee, appointed an audit panel of five individuals to review activities related to certain aspects of the North Carolina school accountability and student assessment programs. Oversight of the panel was provided by the Southern Regional Educational Board (SREB). (The audit panel membership and SREB advisors are shown in Attachment A to this document.)

The audit panel was charged with the review of activities surrounding the May 2001 administration of the North Carolina end-of-grade (EOG) assessment program. The review was necessitated by circumstances that questioned the cut-scores used with the new EOG mathematics tests. The board outlined a series of issues for the audit panel to consider. (These issues- including questions about process and the design of the assessment- are shown in Attachment B to this report.)

This report provides the results of the audit panel's review and makes recommendations for North Carolina decision-makers.

# Background

North Carolina has had a student assessment program for many years, and it has grown over time. The program's design reflects changes that were made as the governor, the General Assembly, and/or the State Board of Education sought to collect more information and to use the assessment results in new and different ways. The program's intent — to gather information that can be useful to decision-makers and can aid in improving education in public schools — has remained the same.

In the mid 1980's, North Carolina began introducing end-of-course tests in the high schools. The tests were administered late in the school term and scored locally. The information was used to assist teachers in assigning final grades and to assist in standardizing course expectations across the state.

In the early 1990's, the assessment program began implementing a similar design for tests to be administered at the end of grades three through eight. These tests were to be designed centrally, and then administered, scored, and reported at the local level. The program's objectives were to establish consistent academic standards across the state and to allow collection of information during the last few days of the school year so teachers, students, and parents would understand how well the academic objectives had been met that year. The test score scales would be linked vertically across the grade levels to permit performance to be tracked as students progressed from grade to grade.

The design's advantages were that the tests were administered at the end of the school year and that the results were available immediately because the processing was done within each district. On the other hand, the design did not allow the DPI to provide timely quality control during score processing, and the DPI would not receive any results until after the results were reported at the local level.

To permit test scoring and reporting at the end of the school term, decisions about the passing cut-scores had to be made and communicated to districts prior to the test administration. In addition, the computer software necessary to permit timely scoring of the tests had to be made available. This design, when combined with the use of scores for promotion decisions and the significant changes in the mathematics curriculum and test items, ultimately contributed to the problems surrounding the May 2001 mathematics cut-scores.

In May 1998, a revised mathematics curriculum was adopted for use in North Carolina schools. New test items were prepared by an outside contractor to match the new curriculum. These test items were reviewed by the DPI, the Technical Outreach for Public Schools (TOPS), and committees of teachers and then assembled into test booklets. In May 2000, the new test items were field-tested by embedding them in operational test forms. The field-test data were analyzed during the summer and fall of 2000; they were used to link the old and new scoring scales and to place the existing cut-scores onto the new scoring scale. For this linking study to produce accurate results, students' relative performance on the new and old items in May 2000 must reflect the relative performance that would occur in May 2001.

This linking design was not ideal, but it met the program requirement that the score linking take place before the May 2001 testing to maintain the tight testing schedules. The DPI planned to conduct another, more rigorous, linking study using data from the May 2001 testing. In fall 2000 and winter 2001, an indication that the May 2000 linking might not be performing correctly was observed, but the DPI continued because no definitive evidence of a problem existed and the department was under pressure to have May 2001 scores produced on schedule. In March 2001, the DPI distributed scoring tables — based on the May 2000 linking — for use with the May 2001 test administration.

When the May 2001 tests were administered, some districts contacted the DPI and reported unusually high passing rates in mathematics. The department informed districts that they should be cautious in using the mathematics cut-scores for the May 2001 test administration and should make promotion decisions based on other information.

The department built a special linking study into the May 2001 test administration. In this study, samples of students took both the old and the new versions of the mathematics tests. Results from this study provided data to link the two tests. In an interim report to Superintendent Michael Ward, the audit panel stated that the DPI approached the linking task correctly, and the analyses should make it possible to compute cut-scores equivalent to those on the earlier test forms. With these results, the school accountability reports could be produced in fall 2001.

The audit panel believes North Carolina has implemented a successful state testing program. Certainly, evidence exists that North Carolina's efforts in the last few years have resulted in educational improvements. However, during the review process, the panel identified several operational and decision-making issues that make the program vulnerable to potential problems. North Carolina policy-makers can improve the testing and accountability program by making changes to the operations and review processes of test development and psychometric analyses and by providing additional resources to support the program. If these matters are addressed properly, the system's integrity, credibility, and fairness will be improved.

The next section of this report presents the audit panel's findings and recommendations.

# Findings and Recommendations

## Findings on the end-of-grade (EOG) assessment program \_\_\_\_\_

Many different ways to design a student assessment program are available for use. Each approach has different combinations of costs, time constraints, features, deliverable products, requirements for psychometric personnel and support contracts, requirements on schools and districts, and implications for students. No single correct design for an assessment program exists. Further, decisions made throughout program implementation can strengthen (or weaken) each design.

From the information gathered by the audit panel, the North Carolina EOG assessment program apparently began as a replacement for the California Achievement Tests used in grades three, six, and eight. When the EOG program initially began, items were developed within the state, department personnel performed functions that typically were contracted out, and test scoring and reporting functions were assigned to individual districts. Previous policies regarding promotion decisions and the California Achievement Test apparently were applied to the EOG program as it was implemented.

As in other states, the North Carolina General Assembly and the State Board of Education increasingly want to use test results as part of various accountability programs and as the criterion for financial incentives for educators and schools. Consequently, the EOG program became a high stakes program where both students and educators had something at risk. In this context, the state must be more concerned about the psychometric quality of the tests, student opportunity to learn the required content, the provision of quality control for all steps of the testing process, maintenance of security for the testing program, and due process issues. The audit panel's view is that the DPI is not well-positioned currently to be responsive to each of these issues with equal effectiveness.

Several aspects of the EOG program are cause for concern:

- Math test items are developed either within the state or by using an outside contractor under very tight timelines.
- The DPI curriculum staff are required to provide content review of test items.
- The DPI subcontracts with a nearby university to format and print hundreds of test forms. The math test forms are produced, printed, and distributed under tight timelines.

- A university contractor conducts psychometric analyses of the tests and provides scoring tables that define the score scales and cut-scores. The DPI distributes the cut-scores to districts before the tests are administered.
- Districts administer the tests, score them, print reports of results, and distribute the results to teachers.
- Districts send test results to the DPI after all the tests have been scored.
- Students who fail to earn a passing score have two opportunities to be re-tested.
- The DPI can not perform complete quality control checks on test scoring and does not have timely access to data. Typically, the DPI will try to review the results from a selected district for an item analysis.

The North Carolina EOG design can be contrasted with a design used in other states such as Florida and Texas:

- Out-of-state contractors develop test items, and committees of practicing teachers within the state review and validate them. Assessment program staff guide this review process.
- The Department of Education curriculum staff participate in meetings related to test design but do not guide or direct content decisions.
- Forms development is contracted out to testing companies that have the expertise and resources to compose, print, and distribute tests quickly.
- Outside contractors are engaged to score the tests and perform psychometric analyses. Department staff members perform quality control checks on the contractor's work.
- Districts administer the tests under security requirements specified by the Department of Education and ship the answer documents back to a central location for processing by a test scoring contractor. Department staff members perform quality control checks on the contractor's work. Results from the analyses are available immediately to the department to identify any malfunctioning items and to allow an advanced review of student, school, and district results before they are released.
- Test reports are professionally designed, prepared, and distributed to school districts by the test support contractor.
- In situations where students may be denied graduation, multiple opportunities are available for re-testing with alternate forms of the test.

The two testing designs are quite different. The design used by North Carolina provides immediate results at the local level, but sufficient checks on the integrity of the system are not inherent in the design. The second design provides quality control

but at a higher cost and less timely reporting of results. The second design reduces the possibility that inaccurate results will be released and used.

The audit panel made the following observations:

The audit panel was told during the meetings that the new mathematics curriculum introduced in 1999-2000 represented “sweeping,” “drastic,” “cutting-edge,” and “revolutionary” changes. The DPI curriculum specialists wanted the new tests to reflect this new curriculum as closely as possible. Such dramatic changes in curriculum and the associated testing are likely to create several psychometric, instructional, and legal problems.

North Carolina law requires that curriculum changes be based on a five-year renewal cycle. Consequently, the assessment program constantly struggles to keep the tests current given the substantial time required to write and field-test new items. Second, new versions of the tests can out-pace the current instruction in the schools, thus denying students an appropriate opportunity to learn the material for which they are being tested. In the 1999-2000 school year, teachers were supposed to teach the old mathematics curriculum (upon which the schools would be graded) while simultaneously beginning to teach the new curriculum. The field-testing of the new items in spring 2000 was based on the assumption that that the new curriculum had been taught.

In May 1999, the DPI contracted with an outside firm to develop approximately 14,000 new test questions in time for field-testing in the May 2000 test forms. The DPI staff members — with the help of various ad hoc committees of district educators — had to review the items and decide which items to include on the May 2000 field-tests. This was a large number of items for the DPI to process in a short period, and the audit panel believes the task was overly ambitious.

The DPI contracted with a local university to prepare test forms. Field-testing the large number of items required about 700 forms to be created in a short period of time. The DPI was unable to get all of the necessary forms shipped to LEAs in time for testing, so not all new items were field-tested in May 2000. The panel was informed that department staff had insufficient time to analyze the statistical data regarding students omitting items at the time the problems were discovered. The audit panel believes that the current arrangement with the university was insufficient to meet the demands of the embedded field-testing system.

The field-test items were placed into the old test forms using decision rules that resulted in an inconsistent positional pattern. That is, the new items were not placed in an identical location within each form nor were they placed consistently at the beginning or end of a section. Since position effect is an important consideration in constructing test forms, analyses of field-test results were confounded.

The audit panel was unable to discover why the statistical information coming from the May 2000 field-tests did not reflect student proficiency accurately, thus incorrectly projected the cut-scores on the new scoring scale. The possibility that some students omitted the items that focused specifically on the new curriculum is one option. The panel was informed that staff members have not had sufficient time to analyze the statistical data with regard to students omitting items found in the test booklets. Another possibility is that teachers in the school year 2000-2001 focused more effectively on the new curriculum, causing students to perform better than the 1999-2000 results indicated that they would.

An independent quality control analysis performed by DPI to check the work of the UNC-Chapel Hill contractor responsible for the EOG linking analysis — based on the May 2000 data — is not evident. However, the panel has no information that would suggest errors were made in the analyses of the May 2000 test results.

The DPI has a mechanism whereby early test data are retrieved from one selected school district, and various analyses and checks are conducted to confirm that local scoring will be done accurately and that all items are performing properly. If problems are revealed, little time is available to verify the conclusions by collecting a larger sample of data or to plan and implement corrective steps. With regard to May 2000, the data-checking process could not reveal the corrections that were needed for the cut-scores.

As early as fall 2000, the UNC-Chapel Hill equating contractor alerted the DPI to the possibility of a problem with the linking of old and new test forms. The North Carolina psychometric Technical Advisory Committee was re-established and met in January 2001 to discuss this and other issues. The committee was informed that delaying the scoring and linking of cut-scores on the May 2001 tests was not possible since these calculations were required for the 2000-2001 student accountability standards. No time, before the release of May 2001 results, was available for the DPI to evaluate those results for problems and, if problems existed, to conduct a new standard-setting study. This situation — attempting to make large changes in teaching and testing within an assessment program that requires both speed and continuity of results — is the quandary of the existing program design.

## A. Recommendations on the end-of-grade (EOG) assessment program

---

1. The DPI needs to develop a plan for addressing intended changes in curriculum expectations so the curriculum changes may be coordinated properly with changes in the tests. Major changes in test content disrupt the continuity of a testing program. Either sweeping changes cannot be made, or if they are made, all those

affected (school personnel, legislators, state government officials, DPI staff, students, parents, and public) must be aware that comparable test results cannot be provided.

2. The DPI should document the various stages of test development — applicable to any test — and adopt those procedures as policy. The legislature and appropriate advisory groups (such as the Technical Advisory Committee) should be consulted to the extent necessary in the development of these policies. A public formulation of state policy puts everyone on notice that the state takes the test development and accountability process seriously and will discourage short-circuiting the procedure in any future test development effort.
3. The test development process for any test should have a set of minimum components including consensus assessment framework development (built on the curriculum standards), item pool development, pilot-testing of the test items, editing and revision of items, building of test forms, field-testing of test forms, scaling and technical analyses of the forms, and operational testing. These efforts need to be coordinated for various subject areas, grade levels, and program components in terms of sequencing activities and meeting timelines. PERT charts that specify the various tasks need to be developed, and the charts should be distributed to all stakeholders.
4. All test development and implementation projects should be accompanied with detailed timelines that reveal the intended and unintended consequences for all major decisions related to meeting or not meeting deadlines. Many times, missing one deadline has adverse effects on several others. All parties need to know as much as possible about the interactions of deadlines.
5. If the current design of the EOG tests — districts provide the scoring services — is to remain in place, then arrangements should be made to obtain a representative data set and review it for accuracy before districts complete the scoring and reporting operation. Appropriate resources should be in place for this review, including the DPI staffing and software support levels and external consultants and experts (such as those on the Technical Advisory Committee). All parties should understand that when unexpected problems occur, the test results may be delayed.

## Findings on general testing and accountability issues \_\_\_\_\_

### **Resources**

In any given year, states have limited financial resources that must be allocated to multiple programs. Education must compete for resources; when the economy is weakening, state departments of education may have to operate with reduced budgets and fewer employees. As resource decisions are made, policy-makers in North Carolina need

to recognize that they have mandated uses for student achievement data that have very high stakes for students and schools. To support these uses, the testing program must be psychometrically sound and defensible. A program that meets these goals must have adequate funding.

The K-12 testing program budget in North Carolina is reportedly between \$11-12 million annually, while the budgets in Texas and Florida are in excess of \$50 million each. While there are differences in design (e.g., Florida and Texas have high school graduation tests and a larger number of students to serve), the financial resources available to the DPI for its assessment operations are inadequate for the tasks at hand.

The audit panel's review indicated that several warnings of inadequate financial resources have been sounded in North Carolina in recent years. In a letter dated July 28, 2000, the chairman and vice chair of the Compliance Commission stated:

“Finally, we also have concerns about the levels of staffing in Accountability Services and, to some extent, in the LEAs. The staffing levels in the DPI's Accountability Services division are inadequate for the scope of work that lies ahead. With the increase in the number of tests for which the department is responsible, the need to update current tests, the number of charter schools, and the changes in the ABCs models, the accuracy of the data will be threatened without additional resources. Mistakes are much more likely to occur when working conditions involve long hours and too few people.”

This call for more resources for testing and accountability was the most jarring, but there were others. The audit panel believes that North Carolina, with the support and confidence of legislative leadership, should conduct a review to determine how additional resources can be provided to support the statewide assessment program.

Moreover, the assessment program has an insufficient number of personnel to complete their assigned tasks. The Department of Public Instruction testing and accountability staff should be commended for their hard work, commitment, and professionalism. The state is fortunate to have a core staff that has done extraordinary work with a high degree of professionalism despite inadequate resources, staff, and support. Hiring people with psychometric expertise is difficult and even commercial test contractors have high staff turnover. State agencies, with salaries lower than commercial companies and school districts, frequently lose trained staff. This turnover is affecting the North Carolina assessment program. Not only has the program suffered staff cut-backs, but the program has lost a number of employees to other positions with greater financial rewards and less stress.

North Carolina policy-makers need to address the issue of insufficient resources. If policy-makers are going to expect the assessment staff to "deliver the goods," then sufficient resources must be made available to do the job. Position reclassifications and higher salaries for mission critical positions as well as additional positions may be necessary. In addition, increased funds for contracts, enabling the DPI to contract out certain functions and use regular staff positions to manage contracts, would be useful. For example, in both Texas and Florida, item development and test validation work tasks have been out-sourced. In Florida, the department gave a contract to a local school district which hired a dozen or more classroom teachers and curriculum specialists to work on the project. One department employee supervises the work of the contractor. The district benefits because its employees are receiving valuable training that will lead to long-term improvements in the district, and the department benefits by having a test development center to work on the test items and test forms.

## **Time and effort**

The audit panel is concerned that policy-makers may lack an understanding of the time and effort needed to produce a testing program that is consistent from year to year and can be used for high stakes decisions. Because of this lack of understanding, schedules are set without sufficient time to guarantee quality tests and to check results carefully before reporting them to the schools and the public. While calendar time seems to be the major variable, person hours is also a factor. To some extent, quality staff of sufficient size can compensate for short timelines, but this is true only up to a point. Many people doing uncoordinated work will not be effective when short timelines are in place.

Informing policy makers about the realities of development and implementation of high stakes tests is part of a solution to the problem. For example, ACT takes two and a half years to produce a set of ACT test forms when no change in test specifications occurs. This timeline includes the time from the initial contract for item writing to the operational administration of test forms. The following is a typical timeline for a state assessment program that engages contractors through competitive bids:

- eight months to complete competitive bidding
- twelve months to create test blueprints, item specifications, and first set of items
- six months to field-test items and analyze field-test data
- seven months to prepare, print, and distribute the test forms to districts
- two to five months to complete all levels of analysis and reporting

Furthermore, the work schedule must be coordinated with the dates for test administration. For example, if test administration occurs during April, field-tests should be conducted in April of the preceding year. Some programs may require field-testing over two previous administrations so enough items are ready for the full-scale administration. Each field-test has to be conducted in the same month as the regular test so the results will reflect the same amount of exposure to the curriculum. Therefore, the implementation of a new program typically requires three or four years of work.

Policy-makers' desire for quick action sometimes prevents them from appreciating the amount of time needed for test development. Compromises made to meet the required schedule can, and often do, return to haunt the program at a later time. North Carolina experienced this in its May 2001 test cycle.

## **Curriculum and assessment**

Tension between the desires of curriculum leaders and psychometricians will always exist. Curriculum leaders, interested in improving the curriculum, want to use large-scale assessment tests to model the direction in which curriculum should change. In contrast, psychometricians are concerned that any changes in the tests can de-stabilize statistical relationships. New, revised tests may no longer be parallel to previous versions and, theoretically, cannot be equated to provide stable score scales and meaningful cut-scores.

The situation in North Carolina is that the tests are being used for high stakes purposes and that students must be given due process. Many current psychometric practices stem from the Debra P. vs. Turlington court case fought over the Florida high school graduation test. One central conclusion in the case was that students have to be given lead time to prepare for the new graduation (promotion) requirement and that the schools must be able to demonstrate that the content has been taught. The implication is that a state cannot change the direction of the curriculum and the state assessment tests overnight.

For continuity in an assessment program, various requirements must be balanced so that the equating process is not damaged seriously and so that schools can adjust their instruction to prevent students from being surprised by unexpected changes in academic demands. With respect to the tests themselves, the major college entrance testing programs — ACT, SAT, MCAT — have been making slight changes to tests for years without serious problems. The key factor is a decision-making process that reviews potential changes to test forms in light of all of the issues — equating, content changes, and form improvement — and then makes decisions by carefully considering a balance among all of the issues. For statewide assessment programs, considering major curriculum shifts on

an eight- to ten-year cycle is appropriate. (The National Assessment of Educational Progress uses a 10-year cycle for revision.) The typical three-year assessment development cycle would provide schools with enough time to adjust their curricula and prepare students for the impending changes.

From a management point of view, assessment staff should be in control with curriculum staff providing support. If litigation occurs, the assessment staff will carry the burden of defense. Therefore, the tests must be designed and implemented to meet psychometric standards (i.e., *Standards for Educational and Psychological Testing*, American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999). Unlike the assessment arena, curriculum and instruction has no parallel professional standards that can be brought to bear in litigation.

## **Standard-setting**

North Carolina assessment operations have used the “contrasting groups” to set cut-scores (i.e., performance expectations) for its tests. This technique is one of several acceptable methods of setting cut-scores, but the panel was asked to consider whether another method would be preferable in the future.

The answer is a matter of judgment as several different approaches (such as the Angoff method or the “book mark” process) can be used. Since the North Carolina assessments consist mostly of multiple-choice items, the methodology chosen should be consistent with that test format. For example, the “body-of-work” method used by the State of Massachusetts would be entirely inappropriate in this case.

If the DPI decides to discontinue the current method, new methods should be pilot-tested carefully with extant data sets. The DPI should provide their standard-setting panels with appropriate training and as much information as possible to ensure well-informed judgments. Further, the DPI should decide in advance what criteria to use in determining the quality and desirability of a new method. For example, if having the “right” percentage of students above a particular cut point is a requirement, this requirement should be stated in advance. If the method’s ease of training and general understandability is a requirement, then that requirement should be stated up front. Methods should be chosen for the pilot-tests using these criteria, and the pilot’s success should be judged using the same criteria. This process should be completed before the DPI implements any new method.

## B. Recommendations on general testing and accountability issues

---

1. As high stakes tests are developed, the DPI needs to consider how motivation of students and teachers interacts with field-test design and implementation of the final version of tests in the schools. Consideration should be given to how initiation of high stakes tests will affect test security and potential cheating, and existing policies (e.g., the Testing Code of Ethics) must be up-dated as necessary.
2. The DPI should review the organization of its curriculum and assessment operations to create a firm chain of command and to assign of responsibilities. Assessment operations and needs should be paramount.
3. The DPI should investigate the possibility of using a “task order” contract. Such contracts are awarded via competitive bids but are executed on a task-by-task basis as the department’s needs change during the test development process. Out-sourcing of test development and implementation functions should not be limited to state universities which may not be capable of meeting necessary timelines or quality standards. For continuity and stability, long-term contracts should be used whenever possible, preferably three to five years in length.
4. Because of declining resources for administering the North Carolina testing programs, substantial staff time is devoted to completing tasks with tight deadlines or to managing crises. Greater resources are needed so increased emphasis can be placed on long-term planning. The DPI assessment staff are competent and dedicated, and they have a laudable “can do” attitude, despite consistently increased demands and decreased resources for the testing program. However, this attitude and their past successful handling of increasingly difficult situations appears to have created unrealistic expectations. The DPI assessment staff need to produce realistic, well-defended plans for the resources they need to conduct successfully the North Carolina testing programs. Funding authorities should provide those resources or reduce the requirements of the testing programs.
5. The legal defensibility of the various North Carolina testing programs (current and anticipated) should be reviewed. This review can be accomplished by convening an advisory committee that includes attorneys who have special knowledge about testing issues and state assessment personnel who have been engaged in such litigation. Including state assessment directors who have been responsible for defending assessment programs in court can be beneficial for describing the kinds of documentation that assisted in the state’s defense.

6. The DPI should initiate plans that provide for quality control operations that are entirely separate from the activities of test support contractors. Numerous instances of errors creeping into state assessment results have been identified around the nation, and no organization is immune from such unintended problems. All statistical and scoring work must be verified independently for accuracy.
7. The DPI staff and the external psychometric consultant were completely responsive in providing information requested by the audit panel. However, much of the detailed technical information was written or provided orally in response to a specific request. That is, detailed written descriptions of some of the technical procedures did not exist before the request. Much program-critical information resides only in the minds of individual experts. In meeting pressing program demands with limited human resources, documentation is often one of the tasks put aside. However, if the program is to be reviewed appropriately and defended legally, the need for complete, routinely produced documentation of all major aspects of the program is critical. Appropriate resources are needed so this documentation can be produced. Also, continuity plans should be developed to protect the testing programs against the losses of experience and historical knowledge that occur with staff turnover.
8. The DPI is to be commended for re-instituting the Technical Advisory Committee. Other states have benefited from input from their TACs, and North Carolina should consider reviewing and increasing the charge to the committee. For example, the TAC input can assist in designing future testing programs, reviewing existing programs, and providing real-time advice on testing issues. The DPI could benefit from allowing its TAC to meet more frequently each year, especially in times when changes are anticipated. The DPI should consider expanding the membership of the TAC to obtain additional input from people who are not state contractors to ensure the provision of independent advice.

Plans should be in place for real-time decision-making during the critical time when test results are being reviewed. Supervisory authorities should be alerted to the fact that the review will take place, be aware of the decisions that they will be making and the implications of these decisions, and be given background information necessary to assist them in making their decisions. In addition, supervisory authorities should be aware of and conform to agreed-upon schedules. The role of different people in this decision-making process should be made explicit (for example, delineate the role that the Technical Advisory Committee has in reviewing results and making recommendations). The greater the stakes involved in the testing program, the greater the burden to check the accuracy and credibility of results before they are used.

## Findings on general decision-making issues ---

North Carolina decided several years ago to begin a concerted effort to improve its educational system. This commitment permeated various levels of government and was felt at the school level in several very specific ways. New programs were created, various advisory groups were constituted, and several innovative student assessments were created. For example, the North Carolina writing assessment program was considered by many to be the model to emulate for many years, and its existence encouraged other states to implement similar programs.

Without question, the North Carolina student assessment and school accountability programs have had an impact on education. North Carolina education has benefited in many ways from this effort and will continue to do so in the future. Current discussions about a high school Exit Exam and the creation of a program to “close the gap” between the performance of majority and minority students are positive steps.

North Carolina’s Legislature and State Board of Education have been active throughout the improvement effort. The Legislature recently asked the Joint Legislative Education Oversight Committee (JLEOC) to study the statewide testing program with the intent of re-focusing the state’s efforts and coordinating the various components at the state and local district levels. At the same time, some degree of confusion and disagreement over objectives and the particular role to be played by the statewide assessments is evident. This confusion can be seen in the wording of Section 28.17.(i) from Senate Bill 1005 calling for a study of:

“Whether the State should consider the use of nationally developed tests as a substitute to State-developed testing.”

“The extent to which additional testing, including field-testing, practice testing, and locally mandated testing, is occurring and whether this should be limited or prohibited.”

And, in Section 28.17.(h), the stipulation that:

“No school should participate in more than two field-tests at any one grade level during a school year unless that school volunteers, through a vote of its school improvement team, to participate in an expanded number of field-tests.”

State testing programs can vary, and the current North Carolina program may or may not meet the state’s needs for the next decade. To review the program and make adjustments as needed is entirely appropriate. In so doing, it is worthwhile to note that Congress is debating the President’s proposals that would demand all states have pro-

grams that routinely measure student knowledge and skills. Decision-makers need information about student achievement, and this situation is unlikely to change in the future.

Program improvements can be achieved by considering how management decisions and processes interact with the needs of the assessment program. With this in mind, several recommendations are made in the following pages.

The following recommendations seek to help North Carolina establish more practical and appropriate test development processes that also meet professional standards. Thoughtful efforts to establish the basic purposes of accountability and to identify the practices and resources necessary to ensure high quality are fundamental for program improvement. Quality test development, standard-setting, and implementation of those standards are crucial to continue raising student achievement.

## C. Recommendations on general decision-making issues \_\_\_\_\_

1. Develop and document processes for test development and standard-setting that ensure technical quality. Align standards of quality with those of the national professional testing organizations found in the *Standards for Educational and Psychological Testing* (1999) published by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education.
2. Review North Carolina statutes and State Board of Education policies to verify that they are coordinated and that they clearly state the purposes of the statewide assessment programs. The design and funding of the testing and accountability programs should reflect those purposes and priorities.
3. Review and revise, if necessary, State Board of Education policies that direct test development and standard-setting operations. Clarify lines of authority and the roles of various advisory committees. Verify that the General Assembly has given specific authority to the board to establish the necessary safeguards and processes to assure the technical quality of testing and standard-setting and the authority to adopt passing scores.
4. Establish a process for State Board of Education and public review of student and school performance standards that includes complete and accurate information, provides adequate time, and involves a variety of stakeholders. Ultimately, the decision-making rests on a fully informed State Board of Education to maintain the balance between challenging, rigorous standards and what is reasonable and fair.

5. Conduct a review of major court cases related to large-scale student testing programs. Determine whether North Carolina's student assessment programs are designed and documented properly so the state can be defended adequately if litigation should occur.
6. Acknowledge operational limitations of state assessments. Make technical quality, reliability, and validity a high priority. Balance the limitations of state assessments with practical and sound approaches to support classroom teaching practices and parental need for information. Continue to communicate clearly the purposes and uses of testing and standards to educators and the public.
7. Schedule periodic, external reviews by psychometric experts to ensure quality test development and standards. Establish a program of research into test development and standard-setting procedures to determine the best method for aligning state content expectations, instruction, and assessment and for promoting continued improvement in testing and accountability.
8. Establish reasonable timetables for test development and standard-setting (allowing adequate time for appropriate planning and technical procedures) to ensure high levels of technical quality.
9. Establish long- and short-range plans for test development and standard-setting. Timetables should take into account the proper sequence and time needed for each task. Program transitions should be anticipated and planned to ensure stability, continuity, and clear understanding of changes.
10. Procedures for major changes in the testing program should be planned carefully at both the state and local levels. Changes in content, testing, and performance standards should be coordinated to ensure the stability and integrity of the entire program are supported. Timetables should reflect periodic, minor changes with less frequent major reviews.
11. Procedures for major changes in the testing program should be planned carefully at both the state and local levels. An appropriate approach is to "phase-in" program changes that refine the operational soundness of the system.
12. The Legislature should clarify the provisions in statute (Section 28.17.(h)) that limit field-testing of tests. All state or national testing programs must be able to gather representative data through field-tests and statistically evaluate new test questions. Without field-tests, no new assessment or any replacement items to refresh previous forms of the test can be developed.

Schools are asked to participate in many types of assessments: district assessments, state assessments, national assessments, international assessments, optional assessments, and special assessments. The national and international assessments (such as the National Assessment of Educational Progress and the Third International Math and Science Study) are conducted on a sampling basis that does not involve many students. However, some schools with particular demographic characteristics may be selected twice within the same year. Sometimes, this situation can be avoided with careful planning.

A statewide assessment program is essential to state policy and should take priority over other types of testing efforts. Even with a state program, the possibility that multiple field-tests could be conducted in a given year does exist. For example, in Florida this school year, students will take field-test items in reading, writing, mathematics, and science. In most cases, the field-test items can be embedded within an operational test, but a separate field-test must be conducted sometimes. The time required in these separate field-tests is usually limited because students take only a few items, not an entire test. However, the information gathered from field-tests is very important in supporting the testing program's technical quality.

## Findings on key groups

---

The groups involved in the planning and decision-making of test development and standard-setting lack a clear definition of roles and responsibilities. The Compliance Commission and the Technical Advisory Committee appear to react to issues as they arise rather than participating in a planning process for test development and standard-setting.

The operation and structure of the Technical Advisory Committee and Compliance Commission should be reviewed and changed. The procedures and relationships are too informal and unstructured. Given the needs of the assessment program, the Technical Advisory Committee meets too infrequently, and the Compliance Commission probably meets too often. The linkage between the Technical Advisory Committee and the Compliance Commission appears to be informal, when the relationship exists at all. The records of Compliance Commission meetings give a mixed picture of focus and engagement with important issues. Since being re-established, the Technical Advisory Committee has met twice (November 2000 and January 2001) in the past 11 months, but no minutes of those meetings exist. These two groups are important to North Carolina — or should be — if they function well and are linked properly to the DPI staff and the SBE.

The DPI appears to have various advisory panels and committees, but the State Board of Education independently can create advisory panels that address functions for which DPI is responsible. This action can create an awkward situation in which the board's advisory panel may offer advice that does not have the quality input that is achieved by going through the DPI advisory processes. Furthermore, advice given independently to the board could work against the proper implementation of a program being administered by the DPI. These situations should be avoided if at all possible.

In addition, the organizational culture may prevent the collection and distribution of full information regarding important issues necessary to the decision-making process. For example, in fall 2001, the technical staff observed that a problem with the cut-scores for the mathematics tests — based on the field-test analyses — might occur. This information was communicated to the Technical Advisory Committee on January 5, 2001. On February 16, 2001, some of the information regarding the cut-scores was shared with the Compliance Commission. The board received a report on this and other related issues on May 29, 2001. This timeline suggests that the State Board of Education was informed inadequately about key technical and operational issues in the mathematics linking.

The State Board of Education sometimes will have to make decisions without the benefit of complete or important information. Therefore, when new and important information is available, a structure and operating “culture” for providing that information to the board should be in place. With the mathematics test problem, the DPI staff and advisors had information that should have been shared, but was not, with the State Board of Education. Perhaps, it was believed that the board would reject any unfavorable and inconclusive technical information that might delay the testing process.

Policy-makers must understand that testing is a technical science as well as an art form. When policy decisions contradict the science of testing, the program is weakened and it may not be able to support policy goals. However, not all decisions can be decided on a statistical basis, so the “art form” aspect of testing involves choices and compromises. In the end, those choices and compromises have to respect the statistical and psychometric science of testing. The State Board of Education must be certain that its policy decisions are always informed by the best technical advice about the science of testing. This means that policy decisions made with “the best information available” must not be viewed by board and staff as irreversible if better or more complete information becomes available.

Test development and standard-setting processes and related decisions ultimately should support the appropriate uses of tests and standards for accountability purposes. Striking the balance between the limitations of state assessments and the need to hold schools and individual students to those standards is a continuing challenge.

## D. Recommendations on key groups

---

1. Regular, independent oversight and review of certain test development and standard-setting aspects should be considered a part of the process. Regular reviews should be established to define the roles and processes for test development and standard-setting and to define significant changes to the standards and accountability program. These reviews should ensure technical quality and operational viability of program changes. Central to establishing these roles is an opportunity for improved local educator involvement, technical review, and identification of key issues.
2. Actions by the State Board of Education should be made based on complete and accurate information with thorough recognition of the consequences of policy decisions. Full, straightforward, and accurate information about test development and standards should be provided to the board to make policy decisions that reflect a respect for the operational feasibility and technical and legal limitations of testing and standards.
3. Establish statutes and State Board of Education policies that assure technical quality, operational viability, and fairness of the testing program. Outline the processes and responsibilities of parties involved in standard-setting and test development.
4. Establish responsibilities for advisory groups to the State Superintendent and the State Board of Education. Those groups should include local educators and various Department of Public Instruction staff. Also included in the responsibilities of advisory groups should be internal and external review of the recommendation making processes.
5. Provide periodic, external reviews of all test development and standard-setting processes. Information from authoritative experts will help maintain the integrity and credibility of the program as well as signal when changes and additional resources may be necessary to improve the program.
6. Establish process standards for testing, standard-setting, and accountability that apply to the Department of Public Instruction staff, the State Superintendent, and the State Board of Education. These standards will provide a basis for making decisions that are independent of internal and external political pressures. The standards will help support agency staff provide professional advice that fully informs policymakers of the impact of their decisions.

## Findings on technical quality and clarity of information \_\_\_\_\_

Inadequate time and inappropriate sequencing of activities did not allow for providing the appropriate, full, and accurate information that is necessary for well-informed decision-making. As a result, the State Board of Education was informed in an inadequate and untimely fashion. However, the short implementation timetable was one of the most important factors contributing to difficulties related to the provision of inadequate information.

## E. Recommendations on technical quality and clarity of information \_\_\_\_\_

1. Establish realistic approaches for compiling and using results from the testing and accountability program. Establish methods that ensure quality, clarity, timeliness, technical accuracy, and appropriate use of the information.
2. Establish priorities for the information that are considered most necessary for the State Board of Education decision-making. Ensure that timelines provide sufficient time for adequate consideration of the issue(s).
3. Develop regular methods for communicating appropriate information to local educators and other state policy-makers including higher education officials, the Governor's Office, and the General Assembly. Educating interested parties with clear, accurate, and timely information is a key factor to sustaining understanding and support of the testing program and standards.

## Findings on oversight of testing and accountability \_\_\_\_\_

Cross-checks and oversight need improvements to ensure the integrity of the testing program. The reliability and validity of testing are extremely important because high stakes are attached to testing results. Equally important is the accuracy of reporting and analysis of results.

Unfortunately, no regular and systematic approach to external review of testing and standard-setting and other important aspects of accountability is in place. Although the current technical consultants provide high quality work, a need exist for additional regular reviews by technical experts.

The establishment of performance standards for individual students and schools are decisions that specifically require more timely involvement of the State Board of Education.

## F. Recommendations on oversight of testing and accountability

---

1. Occasional independent reviews of tests and standard-setting processes would provide support for the integrity and credibility of the testing, standards, and accountability program. These reviews also will provide information for program improvements, substantiate additional resources necessary to improve the quality of the program, and identify areas for school improvement.
2. Policy-makers should have regular opportunities to see how the testing and standard-setting process works. Recent legislation regarding testing and accountability indicates a basic lack of understanding of the testing and accountability program by legislators. Efforts need to be made to help legislators understand how the system can be improved and what resources are necessary to make those improvements

# Align360's

## Additional recommendations on general testing and accountability issues ---

The SBE and the DPI need to create or refine the process for obtaining outside resources so internal talents are utilized more effectively and external consultants are contracted more easily; therefore, work can be completed in a timely fashion and on schedule.

Efforts should be made to recruit and retain quality resources and employees by encouraging them to become invested in the process and by increasing employee motivation. For example, employee motivation can be increase by enabling the sharing and voicing of opinions, showing appreciation for work above and beyond the job requirements, creating an internal recognition system, and trading-off additional earned time to accommodate for periods where overtime hours are necessary. Align360 also recommends exploring the means for creating internal equity in salary and benefits. Turn-over rates are too high to ensure standardization of processes and are higher than most industry standards.

In addition to the panel's recognition of staffing issues, it is important to consider appreciation of current staff verses expectations. One must recognize that the current staffing issues have created a situation where employees are working many hours during evenings and weekends in order to meet very aggressive deadlines. These extra hours need to be acknowledged so that future staffing plans do not include these additional hours as a basis for determining future staffing requirements. Also, appreciation for such professionalism and dedication should be recognized, as employee morale is essential to future work as a team.

The DPI and the SBE should consider streamlining the communication process to reduce the time and steps required to provide information to the State Board of Education. In addition to regular, daily activities, the DPI must spend considerable time preparing information for board meetings, and typically, the time available to complete the process is very limited. Demand for the DPI staff involvement in these activities has increased as the program's focus has shifted to a high stakes program.

## Additional recommendations on general decision-making issues

---

Align360 recommends that the SBE and the DPI create two decision-making processes: Standard and Fast Track. Having these two mechanisms in place for decision-making will alleviate time constraints on minor issues that can be decided quickly or time sensitive decisions that must be made quickly to meet critical deadlines. Guidelines need to be established and publicized for identifying the category for which an issue qualifies. All stake holders must adhere firmly to the detailed process, not allowing an override of the process under any circumstances. A generic decision-making model has been provided by Align360 to serve as a guide for developing the two decision-making tracks. (This model is shown as Attachment C.)

We also recommend creating an internal atmosphere of risk-taking and giving all employees a voice in a forum that is open to all opinions without recourse. Taking on a group responsibility approach for all actions and reactions — not allowing one employee to be responsible for issues or problems — is an important aspect of this approach. The DPI employees should have issues addressed to them as a group, not as individuals.

## Additional recommendations on key groups

---

Align360's additional recommendations As a part of educating policy-makers, Align360 strongly encourages the State Board of Education to tour the facilities available to the DPI. This opportunity will provide an overview of the process for establishing the desired testing program directly involved staff members. Understanding the work that has been done, the process that is followed, the number of staff members working in each specific area, the responsibilities are for each section, and the background work to make this process successful is important. The tour should provide the information and perspective necessary for more effective planning, for creating a more refined decision-making model, and for better allocation of all resources associated with the testing program.

The test development process and the background information on why these steps are necessary should be included in the training for new SBE members.

As mentioned by the audit panel, the roles and responsibilities of various advisory groups need to be clarified. To expand on this recommendation, Align360 recommends that the SBE and the DPI define specific issues such as who appoints members, who sets the agenda, when they meet, what documentation is required (e.g., minutes), and to whom does the advisory groups report.

## Additional recommendations on technical quality and clarity of information ---

Hire an additional person for public relations who will be responsible for all written communication to public, parents, and community. This person would be charged with researching accurate information and eliciting feedback for appropriate response from staff members, and in turn, writing the formal responses with copies to the SBE, the DPI, and any other vested parties. This process will ensure consistency and integrity of responses and group accountability for further action.

# Conclusions

The Audit Panel is impressed with the North Carolina assessment and accountability program and its role to increase the quality of educational programs and student achievement in the state. Apparently, in North Carolina and other states, the support for a comprehensive approach to education improvement including teacher salary increases and school facilities funding would not have occurred without a strong testing and accountability program that showed improvement in student achievement results.

North Carolina policy-makers, especially the State Board of Education, the General Assembly, and the Governor, should be recognized for their support for the variety of approaches, including testing and accountability, to raise the achievement of all students.

The current testing framework of end-of-grade tests in grades 3-8 and end-of-course testing for high school is an appropriate structure for testing and accountability. The approach has contributed significantly to North Carolina gains in student achievement as reflected in the National Assessment of Education Progress and other external measures, as well as being a model for other states.

The setting of mathematics cut scores for end-of-grade tests illustrates that important issues are present and must be addressed to assure the integrity and credibility of the testing and accountability program each and every year.

These primary factors contributed to the recent problems with the mathematics test:

- An implementation timetable that was too short. No time was available for a structured review of results to ensure adequate technical accuracy.
- Inadequate resources and staffing. New tests and new purposes for testing were added faster than resources and staff were added to do the work.
- Major changes were made too frequently to content standards. Significant changes in a short period of time spell trouble in test development and standards setting.

- Inadequate communication to, and direct involvement of, the State Board of Education in setting testing and accountability standards.

The Department of Public Instruction testing and accountability staff is dedicated and hard working. The staff has been over extended in meeting the increased demands on testing and accountability through the 1990s by the State Board of Education, the Governor's office, and the Legislature. The staff turnover rate is an unmistakable signal of serious problems with insufficient staffing and resources that have been cited repeatedly.

The assessment program is not funded adequately, and needs greater coordination between the various interested parties. Design issues are present that contributed to the scoring problem in May 2001. Unless the end-of-grade testing and end-of-course testing program is modified as suggested in this report problems of this nature will likely re-occur.

Assessment programs cast in a framework of high-stakes decisions must be psychometrically and legally defensible for they are invariably challenged on both counts. Although the planned North Carolina graduation test was not a direct topic of study, significant concerns were expressed by the panel about the present plans for the test.

The stakes for the graduation test are even higher than using tests for promotion decisions. This test is likely to be challenged in court, and every effort should be made to assure that the state is legally defensible.

No state approach to testing and accountability is perfect. The use and application of state tests for accountability purposes is a relatively recent development. Maintaining the integrity of the testing and accountability program absolutely requires dedicating sufficient resources, appropriate timelines, adequate oversight, and the involvement of key policy-makers.

Adjustments to the testing and accountability program that ensure quality, integrity, fairness, and practical use of tests will help support credibility and continued support for a model tool for improving student achievement.

# Attachment A

## Audit panel membership

---

Dr. Mary Lynn Bourque, Director, Mid-Atlantic Psychometric Services, Leesburg, Virginia

Dr. Keith Cruse, Senior Director of Student Assessment, Texas Education Agency, Austin, Texas

Dr. Thomas Fisher, Administrator, Assessment & Evaluation Section, Florida Department of Education, Tallahassee, Florida

Dr. Mark Reckase, Professor, Counseling, Educational Psychology, and Special Education, Michigan State University, East Lansing, Michigan

Dr. Wendy M. Yen, Vice President of Research, ETS K-12 Works, Monterey, California

## SREB

---

Mr. Mark Musick, President, Southern Regional Education Board, Atlanta, Georgia

Dr. Jim Watts, Vice President for State Services, Southern Regional Education Board, Atlanta, Georgia

# Attachment B

## North Carolina Department of Public Instruction charge to the audit panel ---

The audit panel is charged to review and comment on decision-making and technical issues regarding the development and processes of North Carolina standards, testing and accountability. The specific focus of the audit is to review issues that influenced recent problems in mathematics standards and testing.

An appropriate timetable to conduct the review would be August through December 15, 2001.

Given the need for timely decisions on school performance and student achievement levels, the review of equating processes and related decisions will be conducted in August.

### **Decision-Making Issues**

1. Examine current state statutes and SBE policies regarding standards, testing and accountability. Specifically review statutes and policies for timetables, reviews and other key factors that frame and influence decisions in those areas.
2. Review the current SBE policy-making process related to standards, testing and accountability and consideration of the impact of decisions. The review of policy development shall include timelines and how they are developed and applied, sequencing of activities, and the review of proposed standards.
3. Examine the standards, testing and accountability review and decision-making role of the State Board of Education. Identify the roles crucial to decisions and comment on the quality, adequacy and clarity of information available to support decisions. Identify specific decision-making factors that led to mathematics end-of-grade test equating problems.
4. Review the role of the Department of Public Instruction staff in decision-making regarding standards, testing and accountability including agency staff interaction with the decision-making and policy role of the State Board of Education.

5. Review the roles and communication of advisory groups including the Technical Advisory Committee, the Compliance Commission for Accountability and other committees that advise the state superintendent on decision-making regarding standards, testing and accountability.
6. Examine the effect of standards, testing and accountability decisions on staff workload, capacity and sequencing of their work. Comment on the adequacy of resources available including staff, budget and time.
7. Examine teacher and public involvement and communication related to standards, testing and accountability, timetables and decision-making. Comment on the quality and timeliness of communication of information regarding changes in standards, testing and accountability to teachers, policy makers and the public.

## **Technical issues**

1. Review the test and standard-setting development processes with a special emphasis on recent changes in mathematics standards and testing. Comment on key elements in the process including the following:
  - Content standards revisions and schedule;
  - Development of test specifications and blue prints;
  - Field-testing, selection of test items and sampling plans, including how schools are selected for field-testing and the mechanics of embedding sample questions;
  - Standard-setting processes including equating and setting of achievement levels for individual students and ABCs school performance calculations (and consequences thereof); and
  - Specific involvement and recommendations regarding math standards, testing and equating of the Compliance Commission for Accountability and Technical Advisory Committee and other advisory groups
2. Review the timetables and sequencing of key events that had an impact on new testing, standard-setting and accountability initiatives. Include in the review content standards and test development review processes and their effect on workload of staff. Comment on how these factors specifically impacted the setting of new mathematics achievement levels.
3. Examine and comment on the planned procedure for equating the new and former mathematics tests and the process for setting achievement levels to be used for the 2000-2001 ABCs Report and for the 2001-2002 school year (i.e. – Student Accountability Standards and ABCs). Identify key factors influencing the quality and accuracy of the equating process.

4. Review current information available for equating decisions relative to determining school performance, including the processes and information used to calculate ABCs performance and bonuses using the equated scores for mathematics for the 2000-2001 school year.
5. Review and comment on the information available for decisions about standards, including comparisons to other standards. Comment on the adequacy, accuracy and quality of that information.

## **Recommendations**

*Comment on and make recommendations regarding:*

1. Improvements to the standard-setting and review process to ensure quality;
2. Processes for equating math scores for use in determining school performance and standard-setting for the 2000-2001 ABCs results and recommendations for mathematics standards for subsequent years;
3. Timetables and sequencing of activities in standards, testing and accountability;
4. The roles and responsibilities of those involved in standards, testing and accountability decisions, including the State Board of Education; Department of Public Instruction, advisory groups, educators and public;
5. Technical quality, clarity and usefulness of information that inform decisions;
6. Processes that help support decisions that reflect appropriate use of tests and standards for accountability purposes;
7. Oversight and review of the standards, testing and accountability processes; and
8. Staffing, funding, time considerations and or requirements to ensure technical quality of standard-setting and testing.

Attachment C



