Part 2

A Review and Analysis of the Tennessee Value-Added Assessment System

Thomas H. Fisher, Ed.D.
Director, Student Assessment Services Section
Florida Department of Education
Tallahassee, Florida

January 1996

(The opinions expressed herein are those of the author and do not necessarily reflect the opinion of any other individual in Florida government.)

EXECUTIVE SUMMARY

In 1992, the Tennessee Legislature enacted the Education Improvement Act (EIA) to restructure various aspects of the state's educational programs. Among its many provisions were several that initiated a new educational accountability program. The accountability program featured a management information system, changes in the student testing programs, extensive public reports, a method of determining improvement goals, and a statistical system to monitor student progress, thus holding systems, schools, and individual teachers accountable.

Certain aspects of the educational accountability program were met with concern and confusion. There was uncertainty as to whether the statistical analyses, known as the Tennessee Value-Added Assessment System (TVAAS), were valid, reliable, and fair. Teachers, particularly, were concerned that they were likely to be evaluated on the basis of a system they did not understand nor trust. In response, the Tennessee Office of Educational Accountability, Comptroller of the Treasury, retained the services of three out-of-state consultants to review and evaluate the TVAAS, especially its utility in evaluating teachers.

This report did not concentrate on the statistical analyses techniques for TVAAS as that was addressed by the other consultants. Instead, the report focused on the design and implementation of the Tennessee state testing program that generates the data for TVAAS and the policy implications of TVAAS as it presently is designed and implemented. The report is based on interviews with various Tennesseans involved with TVAAS, inspection of documents, and limited analyses of relevant statistical data.

The Education Improvement Act and associated administrative rules were inspected. Instead of creating a broad structure for educational assessment and accountability, the Legislature created very specific language that describes the value-added system. This limits the flexibility of those who must implement the laws and modify the program over time. Administrative rules did not appear to have been adopted to support all relevant parts of the student assessment program.

The Department of Education evidently relies upon contractors to do the work needed to support the educational accountability program. Contracts with the University of Tennessee and commercial contractors were inspected. The specifications for the contracts did not appear to be sufficiently detailed to define exactly what the contractors were expected to do. Without more specific language, it would be difficult to hold a contractor accountable for performance. Suggestions were made for improving the bid and contractual processes.

The various components of the Tennessee assessment programs were reviewed. The state uses commercial norm-referenced tests as the basis for the value-added calculations and uses criterion-referenced tests to measure student acquisition of specific

learner objectives. The basic structure of the testing programs is acceptable, although there are areas that can be improved. There are potential difficulties with the continuing need for alternative forms of the norm-referenced tests; problems can arise when the contract is re-bid; national norms may not be a desirable standard of performance; and test administration difficulties may arise at any time such as lapses in test security or schools not testing all eligible students. It was suggested that improvements be made in documentation, program oversight, and implementation of the management information system.

The value-added system was reviewed. It was felt that the system was too complex to be understood by the majority of people. Systems are not in place to verify the accuracy of the processes. At the individual teacher level, the data are dependent on the collection of data from teachers at the end of the year that has not been audited nor is likely to be totally accurate. Value-added calculations at the district and school level are less troublesome than at the individual teacher level. Use of the value-added system at the teacher level was not recommended. Routine release of other student assessment data in addition to the value-added calculations at the system and school level is desirable.

It was suggested that the student assessment system be reviewed to determine how to strengthen the relationships across the various parts and to gather information to monitor student transition to postsecondary education. A master plan should be developed that will better define the data Tennesseans need and the manner in which the data will be collected and disseminated. Overall management and oversight responsibilities need to be strengthened. Tennessee has taken bold steps to improve its schools and its educational accountability system, but there is no single way to foster educational accountability. The search for acceptable and understandable methodology should be seen as an open-ended quest.

Table of Contents

I. Introduction 1

II. Background Information 3

Purpose of Study 3

Study Questions 3

Methodology 3

Limitations of the Study 4

III. Statutory Issues 4

Education Improvement Act 5

Administrative Rules 7

Recommendations 8

IV. Contractual Issues 9

Contract with the University of Tennessee 9

Contract for Writing Assessment with Psychological Corporation 11

Contract for TCAP with CTB/McGraw-Hill 12

Contract with the University of Tennessee for Value-Added Services 13

Recommendations 14

V. Testing Program Issues 16

Overview 16

Exit Tests 16

Purposes of the Tests 17

Design of the Tests 18

TCAP Assessment Tests 18

TCAP Competency Test 20

TCAP Writing Assessment 20

TCAP Subject Area Tests 21

Equivalent Forms 22

Performance Standards 23

Test Administration Procedures 25

Training 25

Test Administration 25

Testing Population 25

Accuracy of Processing 27

Test Security 28

Reporting of Results 30

Management Information System 32

Recommendations 33

VI. Value-Added System Issues 34

Student Proficiency 36

Test Quality 37

Faculty/Student Data 38

Matching Records 38

Data Processing Accuracy 39

Data Interpretations 39

Understanding the TVAAS 41

Evaluation of Teachers 44

Recommendations 47

VII. Conclusions 48

References 50

Appendix A Illustrative Achievement Data by Ethnic Group

A Review and Analysis of the Tennessee Value-Added Assessment System

I. Introduction

In 1992, under the leadership of then-Governor Ned McWherter, the Tennessee Legislature enacted the Education Improvement Act (EIA) designed to restructure education in the state, provide enhanced funding for education, and create an educational accountability program reaching into each school and classroom. EIA immediately affected the educational community, both in positive and negative terms.

EIA had an immediate impact by producing increased school funding, and it changed the manner in which school system superintendents were selected. But, perhaps more visibly, it created a new process to inform parents and citizens about the academic progress of students. The state testing program had been in place for a number of years; however, with the advent of the "value-added concept" schools and systems were now being rated on the degree to which progress goals were being met. These reports were formally presented to the State Board of Education, the Commissioner of Education, and the Legislative leadership and then released to all news media. Newspapers routinely reported the test results. For example, the Murfreesboro Daily News Journal reported the 1995 test results with the headline, "Cannon students do well on Tennessee report card" (Daily News Journal, undated). The Knoxville News-Sentinel headline read "Value-added testing gives Knox schools mark to shoot for despite mixed scores" (News-Sentinel, Oct. 15, 1995). Parents and citizens were able to see the test results for each school, color coded to show which schools were making adequate progress.

The Tennessee Value-Added Assessment System (TVAAS) was scheduled to release results on the basis of individual teachers, but the possibility created immediate controversy. The News Journal headlined one article, "Value-added bad way to evaluate: teacher" (News Journal, undated). The accompanying story described how teachers believed the system was "unproven." Teachers expressed their concerns about being evaluated on the basis of a single test, about the public not understanding the complexities of educating children, and of comparing students to national norms. The president of the Murfreesboro Education Association was rather blunt in her evaluation of TVAAS by stating, "If we had one wish, I think it would be to have value-added eliminated."

In April 1995, the Office of Education Accountability issued a report entitled The Measure of Education: A Review of the Tennessee Value-Added Assessment System (Comptroller of the Treasury, Office of Education Accountability, April 1995). This report was presented to the Legislature and contributed to the general air of concern about the program. It concluded that there were serious questions about the data which could only be answered by experts who understood the statistical techniques being employed. There were unexplained variations in the data which mystified observers and generally led

educators to question the validity of the measures. The report concluded with a call for an independent review of the entire process.

The Education Oversight Committee, which received the report from the Office of Education Accountability, was concerned about the accuracy of the conclusions reached. One state senator stated that "the report was biased against the accountability measures and the value-added model created by statistician William Sanders" (Daily News Journal, undated). The authors of the report denied any bias and repeated their recommendation for an independent study. Dr. William Sanders, the person responsible for the design of the TVAAS, frequently was asked to explain the calculations and sought to explain the process to various individuals and organizations. He responded positively to the suggestion that the program be reviewed. As reported in the newsletter of the Tennessee Education Association, Sanders stated, "I want to suggest that this review be conducted as soon as possible to lay to rest any apprehensions that anyone might have concerning the validity, reliability, and robustness of the entire TVAAS process" (TEA News, January 1995). The Office of Educational Accountability, Comptroller of the Treasury, ultimately decided to conduct such an independent review of the system.

Feelings of apprehension continued to be expressed by educators as the fall data were being released. In commenting on the value-added accountability program, the director of the Nashville Metro School system was quoted in the Nashville Tennessean as saying, "It does not work yet...It most certainly needs the attention to the valid criticism raised" (Tennessean, October 8, 1995). A classroom teacher interviewed in May by the Tennessean was critical of the proposal to use value-added calculations in the evaluation of teacher performance. Said Paul Bass, a science teacher, "I'm confident in my students' ability to test well, but I don't think that's a fair evaluation of a teacher's ability" (Tennessean, May 31, 1995). The Commissioner of Education said she was uncertain of where Tennessee was headed with the value-added system, but "I believe the legislature and the community are entitled to accountability" (Tennessean, Oct. 8, 1995). This theme was echoed by an editorial appearing in the Murfreesboro Daily News Journal. In commenting on the performance of area schools, the editorial stated, "Depending upon how your school system, school or child did on the TCAP or competency test, you may think those tests are either wonderful or badly flawed. But the fact remains, they are the only concrete way to track the success or failure of our efforts to educate our young. Rather than grouse or gloat, we should all be figuring out how to make sure next year's scores are even higher" (Daily News Journal, undated).

In the fall 1995, the Office of the Comptroller issued contracts to three consultants who had no direct ties to the Tennessee Department of Education or to the TVAAS for the review of the value-added program. Recommendations were to be offered to the Legislature by January, 1996, so that changes could be made in the statutes as might be appropriate. This document is the report compiled by one of the three consultants.

While the primary purpose of this study is to consider the efficacy of the value-added procedures, other aspects of the Tennessee testing program must be considered.

The state has embarked on an adventure of great complexity with interdependent components. It is impossible to consider the adequacy of the value-added methodology without also considering the adequacy of the testing program data upon which the value-added process is imposed. The data from the testing program cannot be validated without considering the circumstances under which the tests are administered (e.g., whether security is an issue). The testing program itself might be questioned if the tests were constructed in an inappropriate manner or if the tests had no proven instructional validity. And, from a policy perspective, the statutes, regulations, bidding processes, and contracts must be considered as driving forces for all which follows. One does not just review the value-added system -- all aspects of the entire program must be inspected to determine how the parts contribute to the whole and how changing one component might affect other components. This study will consider all of these issues to the extent possible in the allocated amount of time.

II. Background Information

Purpose of Study

The purpose of this study was to review the Tennessee Value-Added Assessment System to determine its value as a means of evaluating the performance of school systems, schools, and teachers. The result of the study should be a summary description of various aspects of the TVAAS program accompanied by evaluative comments and suggestions for changes where appropriate.

Study Questions

This particular portion of the overall study focused on the "design and implementation of the Tennessee state testing program that generates the data for the TVAAS and the policy implications of TVAAS as it presently is designed and implemented" (Contract, paragraph A. 4). The investigation was to consider what deletions, additions, and/or changes in the existing procedures and policies might be appropriate.

The present report does not concentrate on issues related to the statistical analyses that are the basis for TVAAS calculations nor of the issues surrounding the equating of the individual test forms used in the testing program. These rather sophisticated statistical issues are being reviewed by the other two consultants retained by the State of Tennessee, Drs. Richard Wolfe and Darrell Bock, and will be separately reported. They will review the statistical assumptions made in conjunction with the TVAAS calculations and determine whether those assumptions have been met.

Methodology

The Office of Educational Accountability engaged the services of three out-of-state individuals to conduct the review of the value-added system. The consultants

conducted on-site reviews in Knoxville with University of Tennessee staff involved in the program. Each consultant received packets of information from Tennessee officials including publications, manuals, training materials, letters, memoranda, handouts, newspaper clippings, etc. Achievement test data and TVAAS data were made available through data cassettes and diskettes for computer analysis. These data were manipulated in various ways by each consultant to address different research questions. In addition, the consultants had access to messages posted by educational researchers across the nation who participated in an on-line discussion of TVAAS on the INTERNET through the American Educational Research Association's LISTSERV service.

Limitations of the Study

The Tennessee testing programs and the Tennessee Value-Added Assessment System are complex undertakings. They were developed and implemented over many years and, thus, have hundreds if not thousands of supporting records and documents. The sheer volume of these materials creates the first limitation of this study in that there may be important documents that were not received or were not known to exist. The study reflects those documents which were studied. Conclusions could be affected by unknown facts found in other documents not reviewed.

Secondly, the conclusions of the study are affected by the length of time available to gather information and process it. The testing program and the TVAAS system have been in place for several years and have produced thousands of data records of various types. It is not possible for three consultants to repeat all of these analyses to verify their accuracy. Instead, the consultants inspected selections of data from designated years or samples of system and school data. This is a normal process for auditing complex programs and is usually sufficient to reveal problems. However, it must be recognized that there may be problems lurking in some corners of the programs that were not detected during the relatively short period of time available for the study.

Third, there is no one way to design and conduct a testing and accountability program. The approaches used are a function of the history of the state, the politics of the moment, the resources available, and the capabilities of the organizations responsible for the programs. Different people under different circumstances would create a different accountability program. Therefore, it must be understood that the recommendations contained in this report are those of the author based on his experiences and training. Each conclusion must be evaluated by Tennessee authorities and each recommendation must be weighed in terms of its feasibility and attractiveness for implementation in the Tennessee educational system.

III. Statutory Issues

The underlying principles and objectives of a statewide student achievement assessment (i.e., testing) program are, or should be, established in the laws and

administrative rules of the state. This chapter of the report describes the laws and rules that provide definition to the Tennessee assessment program and value-added system.

Education Improvement Act

The Education Improvement Act, henceforth called the EIA, was approved March 11, 1992, by Governor Ned McWherter. The EIA addressed several different dimensions of public education in Tennessee and was considered to be a significant action strengthening the overall system. It included provisions addressing educational funding, reductions of class size, restructuring the selection of school system superintendents, creation of the value-added accountability program, creation of a management information system within the Department of Education, specification of certain required public reports on the health of the educational system, provisions for the security of the state tests, a requirement for students to pass a competency test prior to graduation, creation of an Office of Education Accountability within the office of the Comptroller with specific functions, and other changes too numerous to recount here. The Governor and Legislature undoubtedly worked hard to write a package that would be attractive and provide needed restructuring to public education. In this respect, the Act can be regarded in positive terms.

In Section 3, changes are specified in Tennessee Code 49-3-340. This provision authorizes monies to be used as incentives for schools that exceed performance standards established by the State Board of Education. The grants are contingent upon establishment of the "data management system." It is not clear whether the mentioned management system is related to the "management information system" described in Code 49-1-2.

On the surface, it would appear to be useful to create an incentive program for schools to improve. However, to fairly administer such a program requires considerable forethought and oversight. At a minimum, it would appear that the Commissioner of Education and State Board of Education would have to do the following:

1. Establish administrative rules to govern the administration of the incentive grant program.
2. Define in rule the focus of the improvement program (i.e., what subject areas, processes, measures, etc.)
3. Determine if schools are to apply for the grants to accomplish certain goals or be funded automatically upon reaching certain pre-defined criteria.
4. Define the criteria by which the funds will be released to schools and the measurement system to be used to evaluate progress toward the goals.
5. Communicate all procedures and criteria to systems and schools so they know what to expect and how it will affect them.
6. Establish procedures to monitor the process to determine if the improvements are lasting and successful in raising achievement over time.

7. Establish and implement procedures to audit the program and determine whether schools are reaching the criteria by undesired means (e.g., not testing certain low-scoring students).

It is not evident from the documents provided for review whether these steps have been taken or whether the program has been fully implemented. While the statute appears to be fairly detailed, administrative rules could not be found which track the law or expand upon it.

The wisdom of a school incentive program can be debated. Several states have adopted programs which provide additional resources to assist students who fail the required competency tests. This follows the Federal Title I model and directs monies to people who perhaps need it most. Alternatively, monies can be used as the "carrot" or incentive to get educators to improve. Incentive programs can be manipulated so the results are not quite what was intended. For example, it is possible for educators to manipulate the scores by selectively choosing students to be tested, by moving attendance boundaries, or by shifting special programs into a school to raise scores. Another approach would be to purchase various test preparation materials known to be tightly tied to the test content. Students thus are "pumped up" for the test but have not necessarily really learned any meaningful content. Tennessee may experience one or more of these phenomena.

In regard to the establishment of the Management Information System, the EIA specifies that the Commissioner is authorized to create the system with the assistance of an Information System Council. This is a very desirable function for any Department of Education. If educators, elected officials, managers, and citizens are to make enlightened decisions about public educational policy, the provision of timely and accurate information is essential. This information would include not only student achievement data but also information about systems, schools, classrooms, teachers, students, resources, demographics, expenditures, curriculum, and other related variables. Such information should be collected electronically if possible and should be routinely reported to various audiences. The data should be consistently defined, collected, and checked for accuracy. More will be said about the MIS concept in a later chapter.

The EIA mentions the state testing program within the context of setting improvement gains and measuring growth toward those goals. There is no specific description of the overall testing program, which led me to wonder whether the testing requirements were described in a separate section of statute. I was directed to Sections 49-1-601 - 49-1-610 in the Tennessee Code. I was not informed of other sections of relevant Code.

If it is indeed correct that there is no other mention in statute of the purpose and structure of the Tennessee student testing program, I believe this is a definite problem. I suggest that the Legislature create a specific section of Code which outlines the Legislative intent, defines the purposes of the testing or assessment program, assigns to

the Commissioner of Education certain responsibilities, gives the State Board of Education specific authority to adopt rules to implement the statute, and defines, broadly, the components of the testing or assessment program. This will serve the worthwhile purpose of informing the public and educators of what is intended and provides the basic elements necessary for a legal defense of the program when it is challenged in court.

The present language of EIA and the Code adopting the relevant passages appears to me to be too highly detailed for statutory purposes. I assume the intent of the Legislature was to define the specific statistical model to be used in the accountability program -- indeed, it is referred to as the Sanders Model after its creator. A preferred approach would have been for the Legislature to write limited conceptual language requiring the Commissioner and State Board of Education to create an accountability program and to report back to the Legislature, in a timely fashion, the details of the proposed system and progress made as it is implemented. By creating language which is restrictively defined, the Commissioner and the Board must return to the Legislature to make any change which time proves to be necessary. This can seriously hamper progress.

As an additional comment, the attempt in paragraph (8) to define the Sanders Model through citations in the literature really does not completely restrict the possible interpretations of the statistical model. Statistical models always have several points at which the researcher must make judgments about tolerances, confidence levels, methods of cleaning up the raw data, which data to eliminate, and so forth. No two researchers, acting independently of each other, would take the same input data and arrive at exactly the same conclusions. So, the language in the law which is well-intended to restrictively define the model does not actually accomplish its objective.

In Section 32 of the EIA, the Legislature defined a requirement for a high school graduation examination to take effect in the 1993-94 school year. The language is not clear as to whether the 1993-94 school year refers to the year in which the test is to be implemented or the year for which the students first must graduate under the requirement. The language does not specify the subject areas to be measured nor does it give the State Board of Education the specific responsibility to adopt the student objectives to be mastered or the passing scores for the test.

EIA includes specific language to remove the state tests from public records access. Taken in conjunction with the earlier language in Section 4 paragraph (9), a strong foundation has been created to protect the integrity of the tests and provides the basis for taking action against persons who steal the test or cheat in some way.

Administrative Rules

In the usual situation, the Legislature enacts statutes and the State Board of Education creates administrative rules to implement the statutes. The rules provide additional detail to the statutory language and provide broader structure wherein the Commissioner of Education can design and implement procedures. The role of

administrative rules should not be underestimated. While states may differ in specifics depending on their legislative and case law history, it seems that the implementation of potentially controversial programs would be enhanced by having a clear definition of authority for policies and procedures enacted by the Commissioner and Department staff.

Tennessee has embarked on a pathway that probably will result in legal challenge at some point. This may occur with the value-added system or it may occur with the student competency requirements. In either case, when litigation occurs, it is of considerable help to those defending the State if the proper laws and administrative rules are in place.

With this conceptual framework in mind, I reviewed the Rules, Regulations and Minimum Standards for the Governance of Tennessee Public Schools, 1994. I found only two places where the state testing programs were mentioned: Section 0520-1-3-.03 (9) and Section 0520-1-3-.06 (d). The first provision lists the grade levels for some (not all) of the tests, provides for storage of the tests, and defines a student appeals period. The second provision provides some detail to the ninth grade proficiency test, defines the passing score, and describes a special education diploma. While these provisions in Rule are appropriate and necessary, they are insufficient to provide the needed definition to the testing program.

I realize my comments regarding the law and statutes have to be taken in the context of Tennessee legal history. However, I have been involved in several legal actions involving Florida's various testing programs. One of them, Debra P. v. Turlington (1979), was a class action case heard in Federal Court which went to the Appeals Court twice and required four years to be settled. Much of the defense of this case rested on clearly defined statutes and administrative rules. Throughout this case and others of a similar nature, a frequent question asked was, "Did the Department of Education exceed its authority as defined in rule and law?"

Recommendations

To summarize this discussion of the EIA and associated rules, I conclude that the Governor and Legislature created a far-reaching and comprehensive education reform package with many good features. One central focus of EIA is the new educational accountability program designed to monitor progress in student achievement over time. My recommendations were

1. Create statutory language which is more general, assigning the Commissioner of Education and the State Board of Education the responsibility to design and implement the programs needed to meet Legislative intent. The Legislature should then exercise oversight responsibilities.

2. Create statutory language which specifically defines the overall structure for the state testing and assessment programs, giving the Commissioner and State Board of Education responsibilities to implement rules and procedures as needed.

3. Create and adopt more detailed administrative rules to govern the implementation of various aspects of the state testing and assessment programs.

IV. Contractual Issues

The Tennessee Comprehensive Assessment Program (TCAP) consists of (1) a series of nationally norm-referenced tests administered in grades 2 - 8; (2) a production Writing Assessment administered in grades 4, 8, and 11; (3) a series of high school subject area tests required by the Education Improvement Act; and (4) a high school level TCAP Proficiency Test, passage of which is required prior to award of a high school diploma. In addition, there are other testing programs including administration of the Scholastic Achievement Test (SAT-I), the American College Test (ACT), Work Keys (from ACT), and the General Education Development Test (GED), a high school equivalency test. Other tests may be administered by local school systems, and some students may participate in the Armed Forces ASVAB testing program. The value-added analysis system is actually separate from the testing programs and is controlled by its own contract and its own staff.

The Department relies heavily upon contracts issued to the University of Tennessee without competitive bidding for services in support of the various testing programs. The University, in turn, subcontracts for certain work and does other work itself. In addition, the Department itself develops and issues contracts for certain test support services. This approach is completely ordinary and is similar to that found in other states. It is efficient and appropriate for the Department to issue a contract to a state university for a variety of support services. From the university perspective, this type of work not only is in the area of public service, but it also provides training for graduate students, provides a ready source of data for research studies, and provides a base of support for test services that can then be extended to meet the needs of the university community and the larger educational and public service arena outside of the university.

With this as an introduction, I now will turn to a review of the contracting procedures based upon the requests for proposals and contracts provided to me.

Contract with the University of Tennessee

The approach used by the Legislature and the Department of Education to obtain the needed testing services depends upon a number of factors including existing law and

budgetary considerations; timelines, and staffing considerations. If the Legislature wanted, a direct appropriation could be given to the University of Tennessee for the testing programs. This approach would bypass all bidding processes and would bypass the offices of the Commissioner of Education and the Department. The monies and responsibilities would flow directly to the University with the expectation that it will design and implement an attractive and accurate program, otherwise ultimately losing the appropriation. This approach has the advantage that the Department would not have to have any staff members employed to work on the projects or to exercise oversight responsibilities. This approach has not been taken, and I do not recommend it. The testing program is too important as an arm of state educational policy to simply be turned over the a state university without advance design and oversight remaining with the Commissioner and the Department.

A second approach would be for the Department of Education to create design specifications and to send them to the University of Tennessee. The University would be invited to submit a technical and cost proposal which would result in a single source contract, assuming negotiations were successful between the two parties. This approach implies that the Department has knowledgeable staff who are firmly in control of the program design and are charged with the responsibility of getting the legislatively mandated program correctly implemented. The Department staff, thus, are in control and the University operates solely to perform the assigned work tasks. Presumably, the University and the Department would operate as a team with final decisions authorized by the Department. Payments would be made to the University in accordance with contract terms and only when work has been performed according to Department standards.

The approach used by the Department in implementing the TCAP appears to be somewhere between the first two models described above. Clearly, the Legislature did not simply allocate monies to the University of Tennessee with the University then making all decisions about the testing program. However, the second model does not appear to be operational either. There is no evidence that the Department issued any specifications for the TCAP. It appears that the contract language is written without any specifications nor any semblance of proposal creation by the University. This raises the issue of the extent to which the Department staff really are in control of the design of the testing programs and exercise oversight and management of the program. The contract, despite its twelve pages of length, is not specific. Only six pages are devoted to the technical dimensions of the testing program, the remaining pages being "boilerplate."

I will not critique the contract in any detail. I observe that it contains only broad categories of duties, many of which go beyond the administration of TCAP (e.g., Title VI, Title I, etc.). To illustrate the breadth and lack of specificity of the contract term, consider the following contractual obligation:

> "The Contractor agrees to perform the following services for the Department of Vocational Education:

a. project consulting, computer programming and processing, and general consulting duties that will facilitate automation of a statewide data collection system.

b. Research analysis required for evaluation of various statewide vocational education programs."

This provision is entirely without any specificity. How much consulting and computer programming? Where? For which programs? What research studies? Which evaluations? When? How extensive? If the University did not deliver the services needed for these provisions and there was a dispute over payment, the Department never would be able to claim that the services had not been provided. This lack of specificity means that the University is in charge of the program, and the Department lacks the tools within the agreement necessary to guarantee completion of the assigned tasks.

The contract is written in the amount of $3,232,260.00 with no notation of a payment schedule nor of any deliverable services that are required for issuance of the payments. It is not clear whether the contract is considered to be "flat fee" or "cost reimbursement." This should be made more explicit. I was not given information to indicate what total amount was paid to the University for the previously completed contract. If it was a flat fee agreement, the total amount should have been paid. If it was cost reimbursement, the total amount paid should be less than the total.

I also reviewed the draft contract between the Department and the University of Tennessee for the next year of services. This draft is not substantially different from the existing one.

Contract for Writing Assessment with Psychological Corporation

The University of Tennessee develops all materials for the Tennessee Writing Assessment Program. Since these tests require students to write responses to specified prompts, as opposed to responding to multiple-choice items, assistance is needed in scoring the great quantity of written material which results. The University prepared a Request for Proposals that was sent to several companies. According to the RFP timeline, the specifications were issued August 26, 1994, and proposals were to be received by September 16, 1994. Allowing time for delivery of the RFP through the mail and assuming overnight delivery of the proposals, apparently the vendors had less than three weeks to respond to the request. In my opinion, this is an insufficient amount of time for a considered response. A minimum of four weeks should be allowed for vendors to respond to such important RFP's.

The RFP I viewed was only five pages long. There was no "boilerplate" associated with the copy I was given; I do not know if such exists. By contrast, the 1994-96 RFP for Florida's writing program was 136 pages long, including 38 for the technical

specifications and 20 pages for the boilerplate clauses. Without the boilerplate clauses, I could not determine how the winning bidder was to be selected.

The proposal received from Psychological Corporation, which was the winning vendor, was lengthy and provided an in-depth response. Unfortunately, this means that the vendor was in the position of specifying to the University the terms of the work agreement. In other words, because the RFP was so general, the University lost its option of telling the vendors what was to be accomplished and how it would be done.

I note that the University consummated its contractual agreement with Psychological Corporation through issuance of purchase orders. For a contractual obligation as complex as implementation of a writing assessment, I am uncomfortable with this degree of brevity. A written contractual agreement would be preferred.

Contract for TCAP with CTB/McGraw-Hill

The only TCAP Request for Proposals available for review was the one issued April 7, 1988, covering assessment activities beginning in the spring of 1990. The RFP required vendors to propose activities supporting administration of a norm-referenced test in grades 2 - 8 and 10, criterion-referenced achievement data in grades 2 - 8, early warning screening of 8th grade students for the high school graduation testing requirement, measurement of students' writing skills, and certain other designated activities.

The technical requirements of this RFP consist of eighteen pages of specifications. The level of detail contained in this RFP is much better than in the previous ones analyzed above. Suggestions for changes are made in the following paragraphs.

The RFP requires the vendor to provide a bid bond and also requires a performance bond. It does not provide specific penalty clauses.

The RFP uses a 100-point evaluation procedure with cost weighted 10 points. There was no discussion of how the proposal evaluation team would rate the required criteria and what the methodology would be for assigning points. This is not a small point since clearly lower cost should generate higher points in the total scheme of things, and this was not explained.

On page thirty of the RFP, the specifications indicate that a one-year contract will be awarded for pilot-testing to take place in 1988-89 with one-year optional renewals for up to five years without further bidding. There is no specified cost proposal outline showing specific worktasks by cost categories. Having this information would be useful to the Department and the University in building invoice schedules and as a basis for contract amendments during the lifetime of the project. For example, the vendor does not have to identify the personnel costs or the travel costs associated with the task of completing the dissemination and training tasks.

The language of the RFP is not clear as to the basis for vendor pricing. (This might have been addressed through correspondence or telephone calls.) The RFP describes basic services for one year and then alludes to the possibility of further amendments covering additional years. The vendor is not told whether to submit costs on the basis of the first year only or to estimate costs for the entire five-year period of time. The RFP does not state whether Tennessee expects the vendors' cost structure to remain the same for the extension periods or if costs can be adjusted. The RFP does not ask the vendors to propose invoice schemes nor does it specify what the payment stream will be.

I was not provided a copy of the contract written to correspond to the 1988 RFP. I could not determine how the RFP language was carried into the contractual obligation.

However, I was provided a copy of the contract between the University of Tennessee and CTB/McGraw-Hill dated July 1, 1994. This contract covered the testing activities for a one-year period of time ending June 30, 1995. This contract format seems strange. Attachment A appears to be the only place where the specific deliverable products and services are articulated. Assuming there was a complete RFP and the vendor replied with a detailed technical and cost proposal, these should have formed the attachments to the contract, not a summary restatement of the deliverables. The RFP and the proposal constitute the sole and complete definition of what has been requested and what has been proposed. The contract should not attempt to restate that agreement. It appears to me as if the language of the RFP is much too thin and does not contain all desired clauses.

Contract with the University of Tennessee for Value-Added Services

The Department of Education entered into a contract with the University of Tennessee for services for the implementation of the "value-added" accountability program. No RFP or specifications were provided for review, and it is not known if any exist.

This contract is very general and does not contain enough information for the Department to control the work of the University on the project. From reading the language, one gets the impression that the University is being given considerable freedom to implement the value-added assessment as it believes to be appropriate. Any given clause of the contract is subject to open-ended interpretation of how it will be accomplished and what will be the deliverable products and services. For example, clause III. A states that research will be conducted to 'identify influences that sustain or impede the academic growth of students." There is no mention of what specific studies will be completed, whether they will be written as reports, how they will be communicated to the Department, or how much they will cost. This begs the question, "How will the Department know when this task has been accomplished?"

The contract clauses mix deliverable products and services with contractor processes. For example, Section III. is titled "Research," so the reader is prompted to

anticipate seeing a list of research obligations of the University. The first clause, previously described, appears to be a deliverable service. The second clause states, "ACT and PLAN data from Tennessee high school students will be used to provide preliminary estimate[s] of sustained academic growth longitudinally among school systems." This is not a deliverable product or service, it is a process or a condition.

This contract otherwise appears to be devoid of detail needed to determine whether the contractor is performing as desired.

Recommendations

My review of the Requests for Proposals and contracts related to the testing and value-added programs leads me to conclude that the State is not exercising sufficient control over the contracting process nor the work of the contractors. While there are elements of adequate contracting processes present, there are several things which should be done. I list these below:

1. For all contractual agreements from the Department to the University of Tennessee or any other university, the desired products and services must be articulated in a set of specifications or RFP even if the University is a single source bidder. If the Department does not have staff members who have the expertise to create these specifications, consultant assistance can be obtained. Moreover, there are many examples of RFP's from other states readily available, and there are many people with this expertise who can assist.

2. The University of Tennessee or any other university responding to a single source RFP from the Department must be required to prepare a technical and cost proposal. The proposal should be signed by an official in the University structure who has the authority to bind the institution to a contract.

3. The Department should structure its contracts with the University so that it contains the contractual boilerplate language with Attachment A being the RFP or specifications and Attachment B being the technical and cost proposal. Additional attachments can be used for any needed supportive materials. The combination of these materials constitutes the formal agreement. Deviations from this would be permitted only through properly executed amendments. The contracts always should include a provision stating that if there are apparent differences between the provisions of the proposal and the requirements of the RFP, the language of the RFP shall prevail.

4. If the Department of Education or the University is issuing an RFP to commercial vendors, the same model should be followed (see #3 above). Prospective contractors should be permitted a minimum of four weeks to respond to any RFP for testing services.

5. The Legislature should enact a law or include proviso language in the appropriations act which permits the Department to enter into contracts which cross fiscal years. In operating a testing program, it is virtually impossible to negotiate any agreement through a competitive process for tests to be administered in the immediate twelve months following a legislative session when appropriations are made. If this proviso language is authorized, the state would not incur any fiscal obligation if funding is not subsequently provided by the Legislature. This would permit the Department to enter into contracts which are more than twelve months in length. For example, it could issue an RFP in October, enter into a contract by January, begin work in February, and complete work activities a full 18 months later, thus crossing three fiscal years. As the testing programs become more complex, this will be a necessary feature.

6. If the University subcontracts with a commercial vendor for any testing services assigned to it through a contract with the Department, the University should be required to follow the same contracting procedures described above.

7. The RFP's written for testing services must be made more specific. Without specificity, the contracting agency will not have the ammunition it needs to exercise contract oversight or to dispute whether a particular product or service has been completed satisfactorily. Moreover, if the RFP contains vague evaluation criteria, a vendor losing the competition will have grounds for disputing the selection of the winning bidder.

8. In any contracts calling for production of materials, there should be language permitting "author's alterations" and "emergency data analyses" up to a fixed dollar amount during the lifetime of the agreement. These provisions would permit the contracting agency to take corrective action quickly and thus prevent delay of the implementation of the testing programs. No formal contract amendments would be needed to expend monies under these clauses since the overall dollar amount of the contract would already have been approved. In addition, in certain, but not all, contracts, it would be advisable to include a fixed dollar amount for consultative services. These monies would be available to the agency to obtain expert advice from psychometricians, curriculum experts, systems analysts, attorneys, or other experts whose wisdom might be needed on short notice to solve a problem with implementation of the assessment

program. Naturally, in the event monies for these purposes are not spent, the contractor would deduct the unspent money from the final invoice.

V. Testing Program Issues

Overview

The Tennessee educational assessment program consists of several different components. These include the following:

1. Administration of the Scholastic Assessment Test (SAT-I) and American College Test (ACT) as high school exit tests as well as for postsecondary admissions purposes.

2. Administration of the American College Testing WORK KEYS program as a high school exit test (also useful for employment counseling).

3. Administration of the TCAP/Competency Test, formerly the Proficiency Test, as a requirement for a high school diploma. This is administered by the State Testing and Evaluation Center.

4. Administration of the Tennessee Comprehensive Assessment Program (TCAP) at several grade levels. This is administered by the State Testing and Evaluation Center.

5. Administration of the TCAP Writing Assessment. This is administered by the State Testing and Evaluation Center.

6. Administration of the Tennessee High School Subject Matter Assessment Tests. These are required by the EIA and are currently under development.

I will comment briefly about the first four of these and then discuss the others in more detail.

Exit Tests

Tennessee faces the same dilemma that other states face when trying to establish a reasonable graduation standard for high school students. Implementation of a minimum skills test, such as the TCAP Competency Test, may challenge the lower achieving students but does nothing to challenge the average and above average students. In an effort to reach these students, the state has established a requirement that students must take one of three so-called "exit tests."

The SAT-I and the ACT, however, are not competency tests. They are broadly constituted college admissions tests, useful in ranking students for admissions purposes

and in monitoring program progress. While these tests are quite well constructed and are respectable for their stated purpose, there is always a danger in using them for other purposes. The most commonly cited difficulty in using the tests is that students self-select the test -- not all students take one or the other, and the population of examinees shifts from one year to the next. Moreover, if the student is taking the test just because it is an exit requirement and really has no motivation to do well because he/she has not made a definite decision to attend postsecondary education, the test scores will not be a true reflection of his/her competency.

The WORK KEYS program is an attractive alternative assessment program recently introduced to the marketplace by ACT. It offers several unique features including a guidance and counseling feature and measures of non-achievement factors such as ability to listen accurately. This assessment has much to offer although its achievement measures in reading, writing, and mathematics are brief. The WORK KEYS assessment is subject to the same dilemma as the SAT-I and the ACT -- how does one get the student to do his/her best if there is no direct incentive to do so?

One way to address this problem is to establish a differentiated diploma tied to performance on one or more of these tests (or some other yet-to-be-defined test). Students would not be denied a diploma but would be able to earn a higher classification of diploma. Alternatively, the diploma could be given an "endorsement" indicating a certain level of proficiency had been achieved. The endorsements also could be tied to a requirement that students take certain courses and earn a designated minimum grade point average in those courses. These approaches would provide an incentive for the students to perform well and would represent achievement above the "minimums."

While I did not spend any time investigating the articulation between the high schools and the postsecondary level, I offer a suggestion for consideration. In Florida, all students entering the community college level are required to take a single college placement test in reading, writing, and mathematics. It has become evident that some 55% of all entering students fail one or more parts of this placement test. This revelation is causing educators in Florida to review the articulation structure between the high school requirements and those of the postsecondary level. The clear conclusion is that many students are entering postsecondary work unprepared to perform adequately. If Tennessee does not now have a system in place to monitor the articulation of its students from one level to the other, it might want to consider doing so.

Purposes of the Tests

The purposes for the exit test have been cited in the previous section and will not be further discussed.

The purposes of the TCAP assessment in grades 2 through 8 are varied, as is so often the case with state testing programs. The tests provide information to the students and their parents about levels of proficiency. The test results are used to inform

educators, citizens, and decision-makers about the levels of performance of aggregations of students at the school, district, and state levels. The value-added program is a statistical function that analyzes the achievement data in a particular manner to provide educational accountability at the school and system levels. The law anticipates using the value-added analysis to create accountability information at the teacher level.

The purpose of the TCAP Competency Test is to provide a minimum competency test requirement for all students leaving high school with a regular diploma. The data also are aggregated to the school, system, and state levels to inform the public and, indirectly, provide for increased educational accountability.

The purpose of the TCAP Writing test is to measure students' ability to write coherently. There are no "high stakes" decisions made with the resultant test scores. The results are made available to the student and his/her parents as well as to the school and system educators. Aggregations of the data are used in the same manner as with the TCAP assessment tests.

The purpose of the Tennessee High School Subject Matter Assessments is to measure student proficiency in specific subject areas. Presumably, the data will be used in the same manner as that from the TCAP Writing test, but, since this program has not been fully implemented, it is not entirely clear how the data will be used.

In general, these assessments function in Tennessee the same way they function in other states with similar programs. The one unique feature is the addition of the controversial "value-added" analysis. People may wish that large-scale testing programs would focus on one purpose, but this is rarely done. Whether we like it or not, student testing programs serve multiple purposes, and sometimes these purposes interfere with each other. This should not be a cause for dismay but should, instead, be a reason for carefully documenting the programs and correctly interpreting the resultant data.

Design of the Tests

TCAP Assessment Tests. The TCAP assessment program for grades two through eight has been redesigned over the last several years. In the 1980's, Tennessee had separate norm-referenced and criterion-referenced testing programs. This structure was criticized as being burdensome to students and schools because of the increasing time demands for test administration and for other reasons that will not be repeated here. A State Testing Advisory Committee recommended in December, 1987, that the program be changed so that only one test would be administered but scores for the two separate purposes could be extracted. This model is being followed with the current TCAP.

To enable a test to perform two separate functions is difficult but not impossible to achieve. The usual approach to building a norm-referenced test is to select items of middle-range difficulty and high discrimination power. These items would be most useful in rank ordering students, thus allowing use of norms tables to interpret the resultant

student scores. A criterion-referenced test, by contrast, begins with a set of approved learner objectives. Test items are created to be good measures of the objectives, without regard to whether the item can be answered by many students or few students. In other words, if an objective is important and students do not perform well, the item may very well be retained on a criterion-referenced test whereas on a norm-referenced test it would be removed. Once student scores are produced by a criterion-referenced examination, they can be rank ordered, thereby producing a norms table. A criterion-referenced test can be interpreted with norms; a norm-referenced test usually does not make a very good criterion-referenced test, however. This is due to the different manner in which the tests are constructed.

In the Tennessee case, the norm-referenced test was selected through a competitive bidding process in which the winning proposal was submitted by CTB/McGraw-Hill Company (henceforth called CTB). Tennessee then built its norm-referenced test around the CTBS/4, a product marketed nationally. The tests measure reading, writing, mathematics, science, and social studies. The initial contract was for one year with options to renew annually thereafter.

To build the criterion-referenced component, CTB added to the CTBS/4 a set of items created by Tennessee for this component. To round out the items, a few items were used from the CTBS/4 (thus serving double duty). As the student takes the test, there is no demarcation between the items, thereby making it impossible for the student to determine which items are in the norm-referenced component and which are in the criterion-referenced component.

Tennessee has two documents that describe the content of the TCAP. One apparently was derived from documents prepared by CTB and is titled Tennessee Comprehensive Assessment Program Norm-Referenced Component (CTBS/4). This is a general discussion of the content of the CTBS/4 written in discourse fashion without any specific list of learner objectives or domains. It is brief, consisting of six pages of text. Teachers would not be able to read this text and decide how to help students perform better by targeting instruction.

The second document is a series of domain descriptions titled Tennessee Comprehensive Assessment Program (TCAP). A subtitle appearing at the bottom of the cover page admonishes the reader to maintain the "TCAP domains and item specifications" in a secure fashion. Therefore, I assume that the proper title for the document is Tennessee Comprehensive Assessment Program (TCAP), Domains and Item Specifications. This document contains clear descriptions of the domains important to TCAP along with subobjectives, specifications for item writers, and sample items. Descriptions are available for Mathematics and Language Arts, including both reading and writing, for grades 1 - 8. These documents appear to have been developed by Tennessee educators as the basis for the TCAP criterion-referenced tests. These are detailed documents and seem to have the appropriate components and design. I did not proofread them nor make any attempt to critically analyze the content.

It is not clear how the criterion-referenced component is assembled and meshed with the norm-referenced test. None of the reports or manuals provided to me describe this process. The technical manuals I had available matched the nationally marketed CTBS/4 but there was no mention of the Tennessee variation. The test administrator manuals did not describe this nor did the explanatory materials used with the educational community. Interviews with Tennessee staff members revealed that the criterion-referenced test items were sent to CTB at the initiation of the contract. CTB made all decisions about the items to be included, their statistical calibration, and creation of alternate test forms over time. There was no indication that CTB has created an overall test blueprint indicating how the tests are to be configured.

I was informed that CTB has an ongoing role in creating new test items for the criterion-referenced test component and that staff members in Nashville have some responsibility for reviewing the proposed items. I was unable to determine these roles and responsibilities, what work was to be accomplished, or who was responsible for accomplishing it. The current CTB contract does not reveal any detailed work requirement for this activity. The 1988 RFP created by the State Testing and Evaluation Center contained several paragraphs of discussion about the use of Tennessee items in the criterion-referenced test component. It is unclear whether there is any other document describing the blueprint for the CRT or if there is any more recent description.

New items for the TCAP/CT are field tested in Tennessee. However, the items are packaged in separate booklets and administered one week after the regular test. The items were not imbedded within the ordinary test booklet. This practice results in performance values for the items that are lower than when the items are imbedded within the regular test booklet. Students do not perform at their best when they can guess that the tests "do not count." Younger students usually are not aware of this, but high school students are all too aware. Whether this is a problem depends upon how the item calibrations from the field test are used later and whether any correction is made for the non-standard test administration.

TCAP Competency Test. This testing program is not considered part of the value-added system, so it will not be further reviewed in this paper.

TCAP Writing Assessment. The Tennessee Writing Assessment Program has been implemented at grades 4, 8, and 11. It is designed to use a focused holistic approach to scoring student essays as opposed to an analytic approach that would yield more detailed information about each student's written product. Each student is given one stimulus prompt and is asked to write an essay that is considered "first draft" quality. There are no pre-writing exercises or attempts to revise and rewrite.

Fourth graders are asked to write a simple narrative. Eighth grade students write an expository piece, and eleventh graders write a persuasive essay. The essays are scored by an external contractor using trained readers who score each paper with a well-defined

six-point scoring rubric. Two readers score each essay with a third reader serving as a judge to resolve discrepant scores. Scores from the two readers are averaged to obtain the examinee's final score. A score of four or above is considered acceptable. No grading marks are placed on the written essay as they are read. A data tape is returned to Knoxville where the scores are reported in various ways to students, parents, and educators. Summary data are provided for public reports.

Several different reports and sets of training materials were obtained and were reviewed. It is apparent that great care has been taken in the design of the writing assessment and that it is consistent with standard practices around the country.

In the ordinary writing assessment, each year's prompts are not rigorously equated to those used in the previous year. The implication of this is that staff can only make general estimations of which topics will perform like their predecessors. If a new topic is slightly harder than the one used last year, the state performance is likely to decrease even if student writing proficiency had increased due to better instruction. I saw no documentation to describe how Tennessee will address this issue.

TCAP Subject Area Tests. The Educational Improvement Act included a requirement that

> By not later than 1993, the development of subject matter tests will be initiated to measure performance of high school students in all academic subjects for which appropriate metrics can be obtained from group administered tests. These tests must reflect the complete range of topics covered within the list of state-approved textbooks for that subject. As soon as valid tests have been developed, the testing of students will be initiated to provide for value-added assessment. Value-added assessment shall be initiated in all academic subjects within secondary schools by 1998-99 school year, and continued annually thereafter. Value-added assessment may be initiated in non-academic subjects at such times as valid tests can be developed which effectively measure performance in such subjects.

The development of subject area tests apparently is being coordinated by the State Department of Education, not the University of Tennessee. A contract was given, after competitive bidding, to CTB for the production of five tests this year.

It is unclear what is expected with the subject area testing initiative. The language of the law does not indicate that student-level competency tests are desired, so it might be assumed that the tests will serve the purpose of school, system, and state level assessment only. No test design papers were shared with me, so I do not know whether the Department of Education or the State Board of Education has an overall plan for subject area tests.

Initially, one might conclude that the development of high school subject area tests is a good thing--indeed, in Florida the same requirement was enacted by the Legislature in the mid-1980's. After the expenditure of considerable effort, the requirement was abandoned. With this in mind, it is worthwhile to consider the implications of the Tennessee requirement.

In Florida's case, a decision was made to create only 25 tests to measure student proficiency in the most frequently taken academic courses in English, science, and mathematics. We entered into agreements with several school districts for the creation and pilot testing of the tests. Each district worked cooperatively with a selection of other districts to create the test blueprints, specifications, and item pools. Half of the pools would be given to all districts in the state while the other half would be maintained in a secure fashion for use by the State. Each pool had from 500 to 600 items. The items retained by the districts had to be maintained in a secure fashion but could be used for their own purposes.

The state assembled its items into usable test forms to be administered within a 50-minute class period. Several were administered during the designated spring testing period, quite late in the year. Problems occurred immediately. Students did not understand why they were taking a state test. Teachers did not understand how the tests should affect grades. Local administrators considered the additional tests to be intrusive. The test results could not be returned to the schools before the end of the school year. There were accusations that the state was trying to create a standardized curriculum. In short, the purpose of the testing program was not clear, and it was not popular at all. The Legislature subsequently removed the requirement.

By contrast, the North Carolina student assessment program has had good fortune with its high school testing program. I conclude that if Tennessee is serious about moving in this direction, it would be wise to survey other states to determine practices that have been successful. Then, if you want to move forward, carefully articulate the purpose of the program and build support among system educators and parents. Tennessee should anticipate that implementation of high school tests will require more financial and human resources. Each test has to be maintained as a separate testing effort with appropriate staffing to enable the program to function. I doubt that present staff can implement such a large program in addition to current duties.

Equivalent Forms. In some school systems and state assessment programs where commercial norm-referenced tests are used, the customers use only one test form per year. When the commercial publisher has only two or three forms available, the customer is limited to rotating these few forms. In other situations, where the state owns the item pool, as in Florida's high school competency testing program, there can be several alternative forms constructed each year.

The purpose of using alternative test forms is to increase test security, lessen the probability that students can copy answers from each other (i.e., if multiple forms are used

in each administration), and prevent answer keys from being created by students and/or teachers. These are considerations in all of Tennessee's programs as well.

I already have addressed the issue of equating in the writing assessment. Here, the problems associated with equating are rather unique. New topics must be used each year and equating of the topics is problematic at best. Statisticians have not developed any method of equating essay topics other than to tie examinee performance to a multiple-choice test administered at the same time as the essay. By combining the scores, a type of "equating" can be achieved. However, this tends to defeat the purpose of administering a separate essay.

In the case of the TCAP assessments, the contractor, CTB, is responsible for preparation of alternate forms each year. In fact, because of the value-added analysis requirements, the EIA specifically required use of "nonredundant test forms" each year. Ordinarily, a commercial test publisher creates only a few test forms to market. These forms, often just two, are considered the shelf product for the lifetime of that particular test structure. The extent to which the contractor can construct additional forms is a function of the number of items available in the pool.

Tennessee has specified that each new form must not overlap with any other by more than 20%. Assume that the first test form contains 50 items. The second form, also with 50 items, could only retain 10 items from the first. Thus, the pool would have to contain 50 + 40 items. The third form, also with 50 items, could only share 10 with either of the preceding two. Thus, it would need 40 new items. The pool would need 50 + 40 + 40 items. This would continue as long as new forms were needed.

Interviews with staff revealed that equating of test forms is the responsibility of CTB, and no oversight is exercised by Tennesseeans. This is not a good practice. Test equating is a procedure in which there are many decisions, not only about initial test content but also about the statistical procedures used. If care is not exercised, the content design will change over time and the equating linkages will drift. It is not known whether an overall test form equating plan has been developed. Moreover, it is not clear where CTB is obtaining the additional test items needed or how long the organization can continue to produce such forms.

The statistical procedures for equating are complex and are being addressed in a separate report by Drs. Bock and Wolfe. Suffice it to say for now that if the forms have not been properly equated, conclusions about student performance will be misleading. In the case of the required TCAP Competency Test, if the forms have not been properly equated, students' rights to fair and equitable procedures are jeopardized.

Performance standards. There are at least three points at which Tennessee assessment programs use "performance standards." First, the statements of learner objectives constitute a type of performance standard. An example would be the expectation that high school students must be able to "calculate area." Second, each

domain measured by the TCAP uses multiple items. For example, four items are used to measure "calculate area." The criterion for mastery of this objective is three out of four which constitutes a performance standard. Third, individual schools and systems might be expected under the value-added system to demonstrate certain gains over time. These also are performance standards. (Note that I am not discussing the term "content standards," an entirely different concept.) The multiple use of the concept of performance standards implies that there may be confusion in any conversation about the Tennessee assessment and accountability programs. Officials should be careful in all written and spoken usage to clearly articulate how the terms are being used so as not to confuse the discussion.

The important question is whether Tennessee has correctly identified and adopted its various performance standards. Apparently, the learner objectives were identified through the efforts of various subject area specialists working with the Department of Education. This is an appropriate way to proceed as long as the State Board of Education ultimately adopts the objectives as state policy. Conversely, the use of criteria for mastery of each objective on the test (i.e., the three out of four rule) is a statistical consideration and probably does not need adoption by the SBE since there is no immediate effect upon the student who does not master a single objective.

The performance standards for school and system performance were created as a direct result of language in the Education Improvement Act mandating them. The language of the EIA ties the improvement performance standards to the value-added analysis system. Quoting from the law, "The goal is for all school districts to have mean gain for each measurable academic subject within each grade greater than or equal to the gain of the national norms." The SBE subsequently approved performance standards (or goals) for academic gain, attendance, dropout rates, promotion, and proficiency.

There are many ways to define standards or improvement goals for schools and systems based upon what constitutes good educational policy. There are no professional standards for setting school and system standards as there are for setting passing standards for high stakes graduation tests. However, as a general rule, I can suggest that school and system standards must be set in a manner which facilitates the kind of improvements desired by the Legislature, the Commissioner of Education, and the State Board of Education. The standards have to be perceived as being reasonable and attainable with effort. If the standards are too low, they will not generate meaningful improvement, and if they are too high, people will view them as being unreasonable. The standards have to be understandable. If they are couched in sophisticated statistical terms, they are likely to generate controversy and suspicion. The standards have to be widely reviewed prior to adoption and widely disseminated after adoption. Educators should be assisted in understanding the standards and should know how they can estimate their standings based on estimated improvements in performance. Last, the performance standards will be useful only if the overall system of data collection remains stable during a given length of time. If there are changes in the tests, for example, the previous criteria are unlikely to remain useful.

In the allowable time for this research study, I could not determine the extent to which these criteria have been met. I am aware that the SBE adopted criteria for systems and a program of incentive awards for schools. I note that the improvement goals are written for the high school graduation test in terms of "90% success for mathematics." It is unclear what this means. Is it that ultimately 90% of a given cohort group will ultimately pass the test? Is it that 90% of the first-time examinees will pass as ninth graders? Since this is a minimum competency test, why would the goal not be to have 100% of the students earn passing scores before graduation?

One has to be very careful when using a norm-referenced test in a high-stakes school improvement model for reasons that will be more fully discussed in the following pages. For now, suffice it to say that the norms associated with a norm-referenced test are useful only to the extent that the students in Tennessee are similar to those in the norm group. If Tennessee students are somehow instructed in a manner to make them different from the norm group, then the norms cannot be used for comparison purposes. Furthermore, to the extent that current norms represent the achievement of a nationally representative group of students whose achievement is generally thought to be insufficient in the first place by many leaders, educators, and parents, use of the national mean as a goal is not going to lead anyone to so-called "world class standards." That is to say, national achievement levels are already inadequate, so why use them as incentive goals?

Test Administration Procedures

Training. To correctly administer a statewide assessment program, a network of cooperating educators is necessary. Tennessee apparently has created such a network with system and school level persons responsible for receiving the test materials, getting the tests administered correctly, shipping the answer sheets, and receiving the resultant data. There is evidence that the University of Tennessee staff conduct training sessions preceding the test administrations and that an annual testing conference is held to further disseminate information about the assessment programs. Training manuals were inspected and appear to be well done.

Test Administration. Without conducting on-site audits, it is not possible to determine whether the assessment tests are correctly administered. The University of Tennessee staff can be responsible for the design of the materials, but the local school staff are the ones responsible for the actual administration. If UT staff do not already conduct random visits to schools as the tests are being administered, they should do so. This always is useful in reviewing the materials and determining how well the system is working. I also suggest a de-briefing conference after the assessments to interview system and school personnel about any perceived and actual problems. This can be done by UT staff through regional meetings.

One constant problem with low-stakes assessment programs such as the TCAP in grades 2 - 8 is student motivation. The value-added system and, indeed, the entire

accountability system is based on the ideal that students perform to the best of their ability. For younger students, this may be a reasonable assumption, but for older students it may not be true at all. As a good indicator of this, consider that when experimental items are placed in a separate test booklet, student performance drops ten to fifteen percent. Even for the competency test, there may not be total student enthusiasm. After all, the test is given to ninth graders. For them, graduation is a long time away. A better time to administer a graduation test would be in the spring of grade ten or the fall of grade eleven.

I understand that Tennessee does not have a system for "pre-slugging" answer sheets with student demographic information. This is a useful procedure and leads to increased accuracy for the answer sheets. While the Department of Education and the Testing Service in Knoxville may not be able to do pre-slugging because the data are not readily available, individual school systems probably are capable of performing the operations. Blank, continuous form answer sheets could be given to the systems and training provided in how the accomplish the task. Then, before the test administration, test runs could be done to demonstrate that pre-slugging will be accurate. If the system makes a mistake in the final operation, it should have to bear the cost of replacement blank forms.

Testing Population. Policy-makers may assume that all students in the relevant grade levels are being tested. After all, if state policy states that the students in grade nine are to be tested for graduation purposes, then it is reasonable to assume that educators will see to it that the students are tested. Unfortunately, it does not work that way.

Educators are concerned about the performance of individual students as well as the performance of students at the school and system level. It is not inconceivable for certain exceptional education, Limited English Proficient, low socioeconomic, or poor-achieving students to be absent on the day of testing or to simply not be included. This has been such a problem for the National Assessment of Educational Progress (NAEP) that the program has taken special steps to increase the number of students in the sample who are actually included on testing day.

There is a more subtle way to avoid having students tested. Local educators have considerable flexibility in determining who is and who is not classified in a designated grade level. In Florida, the number of students who were tested as beginning eleventh grade students in October 1993 was 99,731. When a new and more difficult high school competency test was introduced in October 1994, the number of eleventh grade students decreased unexpectedly to 95,706 only to increase back to 101,513 in October 1995. The same thing may occur in Tennessee with the graduation test or with any of the elementary and middle grades assessments as high stakes are placed on educational improvements. After all, the easiest way to increase one's scores is to reduce the number of low scoring students or to change boundaries or programs to include a number of higher scoring students. No instructional changes need to be introduced if one can gerrymander the enrollment.

I attempted to compare the number of students who were tested with the enrollment data from the Department of Education. I only attempted to perform this analysis with the grade 8 data because of various problems. I had difficulty matching the data files because the UT and Department systems are not perfectly coordinated. I wanted to perform the analyses separately by grade level and ethnic subpopulation but found that no one had data tapes configured in this way. Furthermore, the UT and DOE definitions of ethnic categories are not identical -- one uses a five category system and the other uses a three category system.

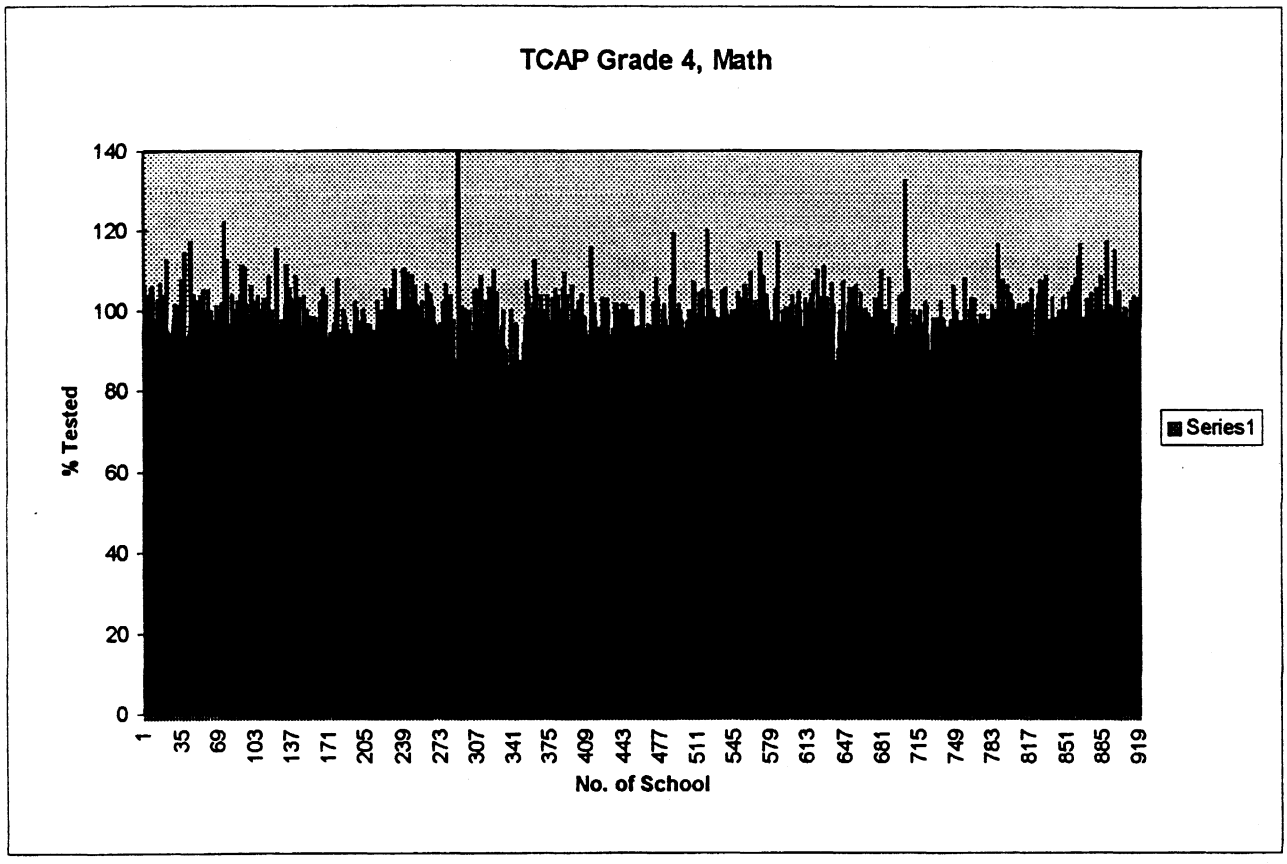
I was able to produce two charts showing the percentage of enrolled students in grades 4 and 8 who were tested with the TCAP in mathematics. The results are shown in Figure 1. As can be seen, there are a few wide fluctuations that can be ignored because they represent unusual situations. However, in looking at the ordinary stream of data, the reader can see that the number of students being tested varies considerably across the various schools. Of course, schools do have fluctuations in their enrollments during the year, so these data alone are not definitive. Only a follow-up audit will reveal whether these anomalies have a natural explanation or whether students were deliberately not tested.

Accuracy of Processing. An assessment program can be properly designed and administered yet result in faulty data. This can occur because of computer programming errors, computer operator carelessness, misunderstandings between the statisticians and the programmers, or inadequate quality control procedures. One should not assume that a commercial vendor is incapable of making mistakes. Indeed, there have been incidents where assessment reports were delayed or where all of the data had to be recalled because of analysis errors. The same can be said of universities acting as assessment contractors: errors will occur.

There are procedural safeguards that can assist. For example, every contractor should be required to have detailed work calendars showing the critical tasks and expected dates of completion. A single person in the agency should be assigned the task of contract monitor to determine if the tasks are being done on schedule and completed with accuracy. Most importantly, prior to release of the assessment data, a team of agency personnel should personally hand check the data, beginning with student answer sheets and progressing through all summary reports. This should constitute a formal data audit and must be designed to reveal any errors in any phase of processing.

In regard to the Tennessee assessments, I do not know what data-checking steps are performed. I saw no evidence that the Department exercises quality control over the work of its subcontractors in the sense described above. If this is not done, it means that each contractor is assumed to be self-checking and capable of auditing its own operations. This is not a safe position to assume.

Figure 1
Percentage of Enrolled Students Who Were Tested



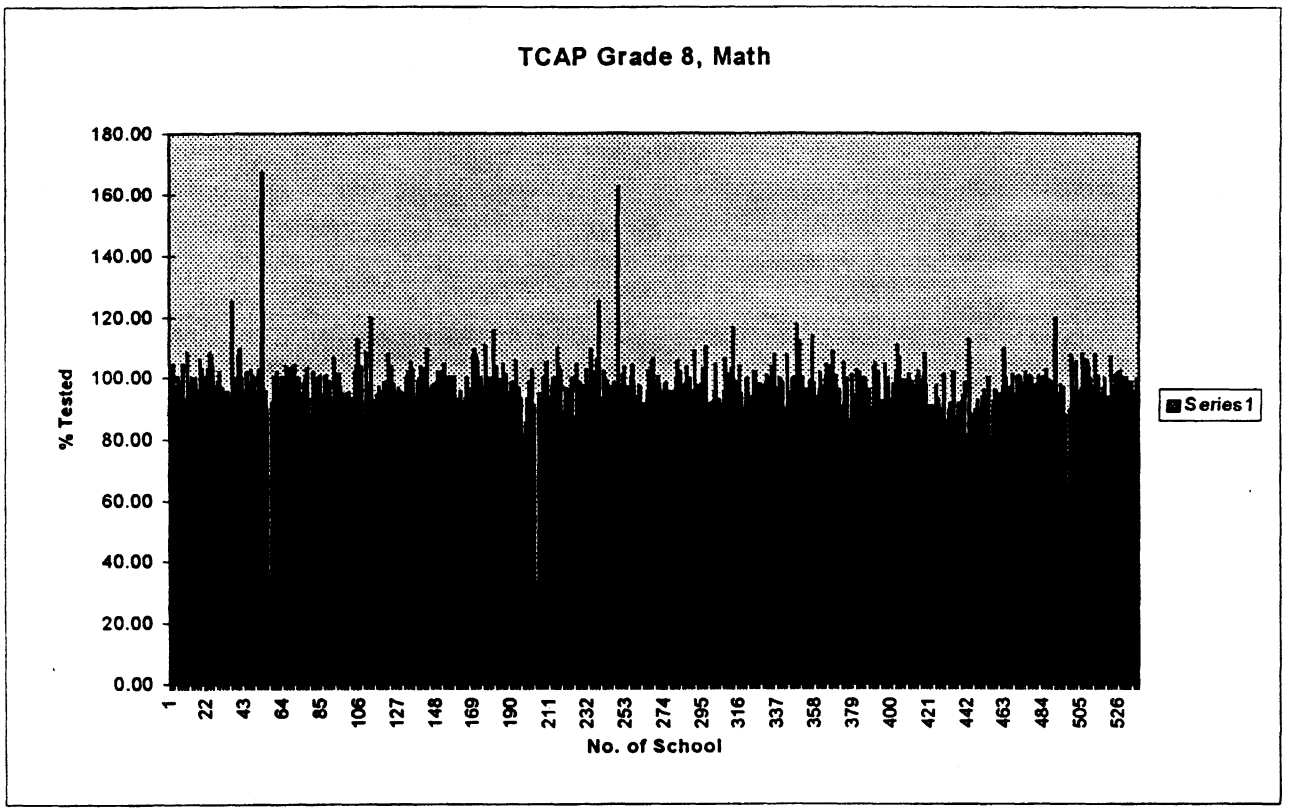
Test Security. Tennesseeans obviously are aware of the test security issue and have taken steps to enforce security. The statutes include provisions prohibiting certain actions such as copying and distributing the tests. Teachers can be dismissed for violating security procedures. Each system is required to develop and adopt test security policies and to submit such policies to the Department of Education.

The policies themselves are clear, although one could always add things to make them more explicit. For example, it is not clear whether teachers are permitted to read/inspect the tests when the materials are located in the schools prior to and immediately after the test administrations. If this is not prohibited, teachers will be free to access the tests and take notes about the questions. Over time, the test items would be compromised and student performance would be increasing for the wrong reasons.

While this study was being conducted, the consultants received a letter from an educator at one of Tennessee's postsecondary institutions. The letter stated that as the TCAP tests were being administered, two sixth grade teachers reviewed the questions and compared the content to the state curriculum guide. They concluded the test content did not match the guide. There was no indication in the letter that the teachers had the authority to access the tests in this manner or that they had been otherwise commissioned

to do such a "study." The Department of Education must consider whether this type of incident is considered a breach of security and, if not, what are the limitations of access to the tests.

Figure 1 (Continued)



It is not clear how violations will be investigated. This is no small problem, because trained investigators may be needed to conduct investigations into security breaches. The assessment staff can hardly be expected to perform these duties. Moreover, an accused teacher, counselor, or test administrator will not want to have rights to due process ignored. Handwriting experts will need to be retained; pencil and erasure marks will be investigated by criminologists. Attorneys will take depositions. I could not determine whether the Department of Education has established procedures for conducting investigations and pursuing security violations.

Without a doubt, there will be security violations and the Department should be prepared for them. These will range from someone stealing a test to test booklets that simply do not get returned. A test administrator may help a student by a wink of the eye or may actually distribute an answer key to students. Educators in charge of the tests may change answers on answer sheets or may actually create answer sheets for students who are absent. Investigating these allegations will require time and courage. I suggest that the Department review its policies and establish a written set of procedures and resources that can be brought to bear on any cheating incident.

Reporting of Results

The Tennessee Comprehensive Assessment Program is an extensive program, as has been summarized in the preceding pages. Many states do not have testing programs implemented at consecutive grade levels and are unable to perform the types of analyses possible in Tennessee.

The data from the TCAP is disseminated to schools by the Assessment Center at the University of Tennessee. Summary reports are available for each school and each school system. Individual student reports are delivered to the schools to be used at the local level for counseling and planning purposes as well as to communicate to parents and to students themselves about strengths and weaknesses in academic proficiencies. The various reports are accompanied by explanatory brochures and diagrams showing how to read the information. The individual student reports are well designed and contain the most interesting information about the student's functioning. Information is available for both the norm-referenced and criterion-referenced components.

The norm-referenced component is reported in terms of standard scores, stanines, and percentile rankings. The results are computer drawn on a scale that incorporates the concept of the standard error of measurement. The criterion-referenced data are reported only in terms of mastery, partial mastery, or non-mastery of the domains. No numerical information is presented, although the results again are reported with a computer generated drawing indicating the concept of the standard error of measurement. The class rosters do not show any numeric information about the criterion-referenced student performance for each domain.

The student report appears to emphasize the norm-referenced aspect of the assessment. The accompanying brochure does not explain the criteria for the domain mastery concept. The "Interpretation of Scores" section of the student report is mostly devoted to the norm-referenced test information. Only one sentence mentions the criterion-referenced information, and it does not explain the derivation of the scores.

I did not find an explanation in any of the Tennessee publications which described how the criterion-referenced domain item sets were equated. In response to my inquiry on this matter, I was provided a brief summary explanation and a lengthier description written by CTB/McGraw-Hill. The document states that the domains are equated by the contractor from year to year but not across grade levels, but there is no elucidation of the technique used. Bayesian estimation processes are used to produce computer domain mastery scores for the student level reports, and this process is described thoroughly. The CTBS Technical Report (CTB, 1991) contains a two-page explanation of the techniques used to estimate domain scores. This explanation is well-written and reasonable. It cautions the user that domain mastery scores are, in a real sense, arbitrary and open to other interpretation.

Inspection of the domain scores over the years 1990 to 1995 revealed a moderate amount of fluctuation, but it is not known whether this variation is due to instruction, population changes, or equating of the domains. The reports to educators, students, and parents did not provide explanations of how the mastery status was determined or if there was fluctuation from year to year in student performance. Furthermore, there were no cautionary notes provided to parents about the uncertainty of the domain estimations.

The official document titled Tennessee Student Test Results, 1994-1995 for the Tennessee Comprehensive Assessment Program is a very nicely produced report, complete with color graphs of various types. The information contained therein is very helpful in summarizing the performance of students statewide. However, I note that there is no information about the criterion-referenced performance of students over time. The only data contained in the report is the cross-sectional performance of students on the criterion-referenced test for the current year. Again, this communicates to the observer that the norm-referenced test is more important than the criterion-referenced test.

The Tennessee Student Test Results, 1994-1995 document includes analyses of population subgroups derived from information about socioeconomic data. Evidently, this information is calculated for school systems based on data available from sources other than directly asking the student survey questions. Other analyses are obtained by asking students to answer certain research questions in the fashion of the National Assessment of Educational Progress. These data are interesting and provide insights into the performance of Tennessee students. Missing from all reports, however, are analyses showing the performance of minority students, exceptional education students, and limited English proficient students.

I also reviewed the 21st Century Schools Program Report Card publication. This is an attractive and impressive report containing information about each school system. It incorporates data on student performance, personnel, expenditures and funding, and so forth. It appears to be a document that will be quite useful to citizens and policy-makers as well as a source of research information for scholars.

The reports I reviewed did not mention Tennessee's participation in the Trial State Assessment conducted by the National Assessment of Educational Progress (NAEP). Certainly, NAEP information is not part of TCAP or TVAAS so its data would not normally be included in reports from those programs. However, in reporting to citizens on the condition of education in Tennessee, the data from NAEP is relevant to understanding how well students are performing. The data from the 1992 Trial State Assessment, for example, is displayed in Figures 2 and 3 below for fourth grade students in mathematics and reading (National Assessment of Educational Progress, 1993).

Figure 2
1992 Trial State Assessment Results for Tennessee,
Fourth Grade Mathematics

	Average Score	% Below Basic Level
Nation	217	41
Southeast	209	52
Tennessee	209	51

Figure 3
1992 Trial State Assessment for Tennessee,
Fourth Grade Reading

	Average Score	% Below Basic Level
Nation	216	43
Southeast	212	48
Tennessee	213	47

The point of this matter is that there is no one way to measure student proficiency in Tennessee. Each different approach provides new information and only by building a web of data can one understand how well Tennessee schools are serving the students.

Management Information System

The Education Improvement Act specifies that Tennessee should establish a Management Information System. Although I made various inquiries, I was unable to determine whether progress was being made to complete this objective. All state departments of education have offices that collect information and store it on computer systems. However, if one has a modern Management Information System, the state's K-12 systems and postsecondary institutions would be connected electronically to permit data and information to be up- and down-loaded to facilitate creation of effective databases and dissemination of needed information. Government officials would be able to query the database and obtain information needed to make timely decisions about program and fiscal policies. I assume this is what EIA anticipates, but I did not obtain and review any document outlining the proposed design of the MIS system.

I did have one simple experience in obtaining information about the K-12 student population. I wanted to match certain records from the testing office in Knoxville with demographic information from the Department of Education. I obtained a data set from each office and attempted to match the records. I found that the data format statements I had been provided were not accurate or complete. After several telephone calls, I obtained the correct formats but then discovered that the two tapes could not immediately be matched because the fields were defined differently with regard to missing values.

More importantly, I discovered after finally matching the data files that I could not answer the questions I originally was pursuing. The files did not contain information about the students by system, by school, by grade, by ethnic classification. Furthermore, the ethnic classification system used by the University of Tennessee did not include the five categories of ethnicity normally required by the Federal Government.

Tennessee has embarked on an adventure into educational accountability which is challenging. Data will be the ammunition used to monitor the educational changes and improvements over time. Without accurate and timely data, the debates will be won on the basis of personal credibility and folk knowledge. I recommend that the Management Information System be given high priority and sufficient staff and resources to allow it to be fully functional. Citizens will benefit from access to the information provided by the system.

Recommendations

Tennessee has taken initiatives to design and implement a comprehensive student achievement monitoring system. Students are tested at several grade levels in the core subjects using both commercially available norm-referenced tests and locally constructed criterion-referenced tests. A high stakes high school graduation testing requirement has been implemented. However, there are places where the overall structure can be improved. The previous discussion contains many individual recommendations, but, in addition, I have prepared in the list below several general recommendations pertaining to the student assessment program:

1. The role to be played by the exit tests and the subject area tests is not clear. I suggest that these be reconsidered in terms of an overall educational accountability program to make certain they will serve Tennessee's needs. In my opinion, the state would be better served to concentrate on exit requirements and postsecondary articulation issues rather than high school subject area tests.

2. Administration of norm-referenced tests at every grade level may not be necessary to accomplish the overall goals of the state. Money can be saved by testing students at two or three selected grade levels. If districts wish to pay the cost of testing students at the "in between" levels, they can do so.

3. The test development and assembly process is not well documented. The state should be in total control of the testing programs and should take steps to create and implement test specifications for all operations.

4. The State Board of Education should review all existing rules pertaining to assessment issues and should then create and adopt any rules that are needed. These rules should address the objectives to be tested, the responsibilities of the systems and schools to teach the objectives, details

concerning the implementation of the assessment tests (e.g., accommodations for handicapped or disabled students), and necessary passing scores.

5. On-site audits of test administration processes at the local level should be conducted by Department and contractor staff.

6. Efforts must be made to monitor whether all eligible students are being tested.

7. Audit procedures should be established by the Department of Education to verify the accuracy of processing by all assessment contractors.

8. The Department should establish clear communications with schools concerning maintenance of test security. Procedures and policies should be created to permit the Department staff to initiate investigations of security infringements and to gather such evidence as is necessary to present in court proceedings (e.g., handwriting evidence, fingerprint documentation, or erasure markings on answer sheets).

9. I recommend that Tennessee routinely participate in the NAEP Trial State Assessment and report the results to its citizens along with other information about student achievement.

10. The Management Information System should be strengthened as soon as possible. Legislators, citizens, and parents want and need timely information about the schools. I recommend that the MIS functions for K-12 and postsecondary levels be coordinated within one administrative structure.

11. I recommend that the state prepare analyses of the student assessment results that include demographic subpopulation performance information. While this information at the school level will be too voluminous to provide in wide circulation, it should be routinely be made available to local school systems and to decision-makers who need it.

VI. Value-Added System Issues

Section (2)(A) of the Education Improvement Act specifies that Tennessee will adopt performance goals for each public school which shall "include but not be limited to determinations based on the current status of each local school system determined through the value added assessment provided for in subsection (g) of this section." The goal for each school system would be to have mean gains greater than those of the national norms. Section (g) then define the "value-added" assessment system as a statistical model to account for differences in prior student achievement so that the impact of teachers,

schools, and systems could be estimated. This impact is defined as the "effect" of the teachers, schools, and systems on the progress of students.

For many years, researchers have attempted to understand the relationships between variables available within schools and systems. Various approaches have been used as researchers gained access to larger, better-defined databases and as statistical models improved with the advent of more powerful computers. Some of the more well-known studies attempting to understand the relationships among educational variables are the Equality of Educational Opportunity studies in the 1960's (Coleman, et al, 1966), the effective schools research described by Brookover, et al (1978), the National Assessment of Educational Progress, and the National Educational Longitudinal Studies (Green, et al, 1995).

One typical method of analyzing school phenomenon has been to collect achievement data and define it as an "output" variable. Then, various other variables are collected and referred to as "input" variables -- e.g., family income, educational level of parents, etc. By using a regression approach, one can analyze the relationships among these variables across large sets of student or school data to create prediction models or analytic models to understand how the variables relate to each other. Typically, the conclusions of such studies discuss the percentage of variation in student achievement scores that is accounted by the input variables. It has proven to be impossible to account for all of the variation in scores because not all relevant and important variables have been measured (i.e., the "specification problem") and because there are always errors of measurement and estimation involved in any regression approach.

In some states (including Florida), regression models have been used to create "predicted scores" for each school. Then, the school's observed (current) score is compared to its "predicted" score as a measure of educational accountability. The school may exceed or may fall short of its "prediction." This approach fell into disuse with the advent of minimum competency programs in which all students were expected to meet the minimum. This was incompatible with the concept of a predicted score.

The value-added model developed by Dr. William Sanders at the University of Tennessee is founded on the "Henderson mixed-model methodology." This is a very complex statistical methodology that fundamentally is a repeated measure, regression system that makes use of all available information, not being negatively affected by missing values in the data set. The label "mixed model" is derived from the fact that a score is a sum of both fixed and random effects (Hays, 1973, p. 551). In the Tennessee situation, teacher effects are considered to be random while school, program, etc. are considered to be fixed. The model is able to estimate the effects of a teacher, school, or district without actually having the "gain score" as one of the data entries.

No attempts will be made here to explain how the value-added system is calculated as this is beyond the scope of this paper. The value-added system collects standard scores from the student achievement test in five content areas: language, reading, mathematics,

science, and social studies. The student scores are collected over a three-year period. If student scores are missing for one or more years, this does not affect the calculations. (Missing values are not "imputed" to complete the missing cell.) If a school closed or a new school has opened, some data will be missing, but this does not affect the analysis. By using each student's previous year's data, the student presumably provides his own control, thus eliminating the need for measures of other factors affecting the student's environment (Sanders and McLean, no date).

One controversial aspect of the value-added model is that information is collected on each teacher in grades 3 - 8 who had contact with each student. The data are then manipulated to produce teacher "effects" that feed into the individual teacher's performance evaluation. The teacher-level data are to be made available to the principal, superintendent, and school board members. This feature of the value-added system has not yet been implemented.

The Sanders approach is contingent upon several assumptions and some very specific data. Some of these are listed below.

1. The system assumes that the test data represents an accurate measure of students' proficiency.

2. The system depends upon tests that are psychometrically respectable in terms of reliability, validity, freedom from bias, correct administration, and proper equating.

3. The system depends upon the collection of accurate information from teachers that describes their contact time with each student during the year.

4. The system depends upon accurate matching of student records over time.

5. The system depends upon accurate processing of the data at the University of Tennessee.

6. The system depends upon meeting various statistical assumptions related to the mixed model itself.

Each of these factors will be described and commented upon in the following paragraphs.

Student Proficiency. The value-added system assumes that the student scores are an accurate reflection of the individual student's academic proficiency. The difficulty with this assumption is that it may not consistently be true. Students who are young may not be suspicious or cynical about the testing situation, but they may be negatively affected by the psychological pressure of performing on a multiple-choice test. As students get older, they become more cynical and more likely to just wander through the test, either marking answers at random or not exerting their full effort.

For individual students, lack of sincerity in taking the test will lead to an underestimation of that student's proficiency and will negatively affect the mean performance for the school and the teacher, if teacher effects are being calculated. The student may have had a good attitude the previous year and may have a good attitude the next year; one cannot be certain what will happen.

In an individual classroom or school, there can be a significant positive effect simply by having teachers and the principal take steps to increase student attention to task. Some schools go to great lengths to accomplish this -- pep rallies, breakfasts on test day, practice sessions days before the test, modifications in the testing setting such as turning off the school bell, and so forth. These techniques will raise scores even if learning has not improved at all.

Schools also can use various "score high" techniques and materials. In the most extreme case, people may have access to an unauthorized copy of last year's test. In the more common case, educators have purchased various practice materials readily available on the open market. It is difficult to evaluate whether use of these materials constitutes legitimate, focused teaching or improper "test prep" materials.

From a philosophical point of view, it can be argued that teachers should not concentrate on the content covered by the norm-referenced test since it is designed to be a general measure of broad content. To the extent that Tennesseans concentrate on the same content, comparisons to the norm group (which did not have the same practice) would be invalid. By contrast, teachers should be teaching the skills measured in the criterion-referenced test since these are skills thought to be important by Tennessee policy makers. This matter gets very confusing when one test contains both sets of materials.

Test Quality. Previous discussions in this paper have addressed the general content and design of the TCAP examinations. I cannot find fault with the content design, the test construction, quality of the items, or test administration techniques. However, I must call attention to the fact that the science and social studies tests are rather short. I would recommend increasing their length to 40 items each if they are to be used for anything other than system reporting. Test form equating is a continuing challenge and is likely to get worse as more forms are constructed from the same pools. The adequacy of the present test form equating is not addressed in this paper.

It is not clear whether Tennessee has planned for the expiration of the current test series with CTB/McGraw-Hill. The present system is driven in many ways by the administration of a particular test series administered in sequential grade levels. When the contract is bid again and perhaps won by a different contractor, adjustments must be made to permit continuity in the accountability system. I did not see any evidence that such planning has taken place.

Faculty/Student Data. The value-added system is dependent upon collection of information from faculty at the end of each school year. These data are collected through the "Faculty/Student Data Form," an optically scannable form to be completed by each teacher. Each teacher places his/her name and social security number on this form and then codes information for each student in the classroom. The Tennessee Department of Education understands the complexity of this form and the potential for controversy associated with its use. Staff prepared a helpful training video to teach classroom teachers how to complete the form.

The information gathered through this form is supposed to be an accounting of whether the student was taught reading, language, mathematics, science, social studies, or all of the subject areas by that individual teacher. Under the terms of the value-added system, a student must be taught by an individual teacher for 75 days per semester to be "claimed." This is based on the concept that a teacher should not be evaluated on the basis of students who were with that teacher for less than 75 days in the semester.

The form requires the teacher to first determine which students have been in attendance at least 75 days in either semester. This is to be done either through attendance data or "personal knowledge." Then, the teacher must estimate the percentage of instructional time spent with each student. For self-contained classrooms, the teacher would be responsible for 100% of the student's instruction. If the teacher is involved in team teaching or if the school uses departments with students moving between teachers, then the teacher must estimate the percentage to the nearest ten percent.

There evidently has never been an audit to determine the accuracy of coding on the Faculty/Student Data form. This raises the specter of difficulties in the analysis if the data are not accurate. It is easy to see how the data could be inaccurate -- completed at the end of the school year, no auditing of the information, the possibility of teacher irritation at having to complete the form, etc. The teacher, in claiming the individual student, is placed in a position of being responsible for the student even if the student has not accepted his/her responsibility for trying to learn. If a teacher has been assigned a student teacher for the semester(s), the teacher is responsible for the education even if the student teacher did the actual teaching.

Matching Records. The value-added model depends upon having matched student records across time. In a perfect situation, the student records would be identified through a control number like the social security number. Each student then would have a dependable and consistent identifier. While many, if not most, students have a social security number, the correct number may not be placed on the test answer form or on the Faculty/Student Data form. Students, especially if young, may not remember their number accurately. Or, more likely, students and/or teachers will not properly grid the number. (Note that students cannot be forced to use a social security number.) In Florida, we have discovered that this is a significant problem even with the college student testing program.

According to information I was given, the value-added staff at UT do not contact the school or the system to verify or correct any social security numbers. Instead, the system depends on a complex operation in which various bits of identification information on the student (e.g., name, birth date, sex, race, etc.) are used to eliminate non-matches among number conflicts (Schneider, undated). Staff assert that they have achieved a 90% match rate using this system.

For school and system analyses, a 90% match rate might be adequate except for small schools. However, for the calculation of teacher effects, it becomes a more significant problem if an individual teacher or school has a large number of mismatched students. I had no information to determine the extent to which mismatches vary across teachers, schools, and districts.

Data Processing Accuracy. I could not perform checks on the accuracy of data processing on either the TCAP processing or on the calculations of the value-added system. It is apparent from conversations that there is no other external check routinely performed on the processing of either component. However, the Tennessee Division of State Audit has conducted an on-site visitation to the value-added system offices to gather information. Dr. William Sanders invited another researcher to do a review of the data analyses, and the findings of that limited study were favorable (Stroup, undated).

My experiences both in Florida and Michigan have impressed me with the absolute importance of having all data checked for accuracy, even if the contractor is a respected commercial firm or a state university. Mistakes are made by the best of us, and accuracy is maintained only when external parties can double-check the accuracy of data. If the Department does not have personnel to perform such checking, a subcontractor can be retained to do the work -- a commercial accounting firm or educational measurement company would be appropriate. Alternatively, the Department can ask for the assistance of a commercial vendor in establishing a protocol for data checking. Then, after training, Department staff can do the work. In the case of the value-added process, the system complexity makes the task of data checking formidable but not impossible.

Data Interpretations. In reviewing the various publications containing data from the Tennessee assessment program, I noticed that emphasis was placed on the value-added data. The system-level information, reported in terms of "percent of national gain," was highly prominent with no accompanying information describing the average scale scores nor the performance ranges of students. I did not see any reports with school level information, although it appears that such is made available to the individual schools for their use.

By focusing only on the value-added information, observers can miss important dimensions about educational levels. To demonstrate this point, I pulled a sample of schools that had obtained the highest relative gain scores in total mathematics. (Note that the same thing could have been done with any subtest.) By reviewing these data, I could immediately see that despite the fact that these were the schools with the highest gain

scores (1) there are some schools with as few as one student in a grade level, (2) the number of minority students varies across schools, (3) the obtained average school scores vary across schools, and (4) the obtained average scores vary across minority groups.

If one takes example schools and further breaks the data down to the student level, it is possible to understand the range of student performance within a school that has been classified as having very high gain scores. These data are displayed in Appendix A, by the three minority groups for two schools (names withheld). It can readily be seen that there are differences between the minority groups, and there is a large range of performance difference between students. To appreciate these differences, one can refer to the student percentile tables provided by the test publisher. Tables 1 and 2 below reveal the student-level national percentile associated with the various score levels for total mathematics.

As can be seen, even though these schools ranked among the very best with high gain (value-added) scores, the average student performance as measured against the national percentile rankings was moderate at best -- the 54th and the 41st percentiles, respectively. Moreover, the performance of minority students was below that of white students in the two sample schools. And, most importantly, from the data in Appendix A, the students ranged in performance between the extremes captured in Table 3 below.

The implication of these data is that when one looks only at gain scores, it is possible for a school to have impressive gains yet have overall modest performance within the school. It also is clear that within any given school there are students who are performing quite well in addition to those who are not functioning very well at all. The TVAAS calculations and the public reports make it possible for schools and systems to be complimented for gains despite the fact that there are many students with serious educational needs within the school population. Citizens and parents can be misinformed about the nature of the strengths and needs of their local schools in such cases.

Table 1
System Orange, School 1
Grade 8, Total Mathematics Scores

	All	White	Black	Other
Scale Score	780.57	798.0	767.60	--
Student %-tile	54	70	41	--
Number	166	69	96	--

Table 2
System Apple, School 1
Grade 4, Total Mathematics

	All	White	Black	Other
Scale Score	692.39	730.6	677.6	678.6
Student %-tile	41	80	28	29
Number	85	23	39	23

Table 3
Range of Student Performance
in Two Illustrative Schools

Grade 8, System Orange School 1	White	Black	Other
Scale Score Range	693-888	603-869	--
Percent below 30%-tile	15	29	--
Grade 4, System Apple School 1			
Scale Score Range	649-781	530-758	564-767
Percent below 30%-tile	8	42	

Understanding the TVAAS. In the opening discussion of this paper, I cited various news articles concerning the Tennessee Value-Added Assessment System. The common theme of the comments from educators was that the value-added system was not readily understood nor trusted. This is a critically important matter and may dictate whether any complex analysis of this type can survive politically.

The value-added system has potential for controversy for several reasons. First, it was created by the Governor and the Legislature as opposed to being jointly developed with the educational community. Second, it seeks to evaluate systems, schools, and teachers and hold them accountable in a very public manner. Sanctions can be imposed if performance is not adequate. Third, it is derived from and dependent upon use of nationally norm-referenced tests. Fourth, it is a very complex statistical model destined to be misunderstood by the vast majority of people.

In regard to the first two items on the above list, I will only state that there are many historic and political reasons why the Legislature desires a program of educational accountability. I strongly support the need for educational accountability programs and will not question the Legislature's responsibility to adopt a program of its own design. However, I will offer comments regarding the third and fourth issues above.

Educational testing is a hotly debated topic in professional education circles today. Among major educational organizations, norm-referenced testing programs are often viewed as being out of touch with current theories of how students learn. This has created a demand for use of so-called "alternative assessments" and "performance tasks" rather than multiple-choice questions. This can be seen in the highly visible new Kentucky assessment system and in the attention given to the Vermont writing portfolio system. The implication of this is that the educational community is likely to suggest that the TCAP, with its reliance on norm-referenced tests, is not moving Tennessee in new directions. Of course, TCAP also has a criterion-referenced testing component measuring local objectives; however, other than the writing assessment program, there is nothing to suggest that TCAP is moving to include more modern measurement approaches. Unless this is addressed in plans for future assessments, it is unlikely that the educational community will be totally satisfied with the current measurement system. On December 8, 1995, the Tennessee State Board of Education adopted a resolution calling for plans for a new student assessment policy. This presents the opportunity to organize new thinking about the state testing programs and is a step in the right direction.

It seems clear that the average citizen, educator, or legislator does not understand the calculations behind the value-added system. The Education Improvement Act states that the value-added model will be

> mixed model methodologies which provided for best linear unbiased prediction or similar language setting forth the methodology used for evaluating measured progress of students, teachers, schools, or school districts, [having] the meaning and be interpreted as set forth in the following references...

This sentence was followed by a list of six professional journal articles. Therefore, the definition in law is dependent upon cross references to statistical journals. In referring to the first of these, I found the following definition:

> "The mixed linear model is given in matrix form, using the notation of boldfaced print for all matrices. Henderson's notation and results yield the following:
>
> $$y = \mathbf{XB} + \mathbf{ZU} + \in$$
>
> where **y** is a m X 1 vector of measured responses, **X** is an m X p known matrix with the rank of **X** ≤ m, p, **B** is a p X 1 vector of fixed effects that are unknown, **Z**

is an m X q incidence matrix, **U** is a q X 1 vector of random effects with E(**U**) = 0, $\in$ is an m X 1 vector with E($\in$) = 0, and

$$\text{var}\left[\frac{U}{\in}\right]=\begin{bmatrix}G0\\0R\end{bmatrix}$$

In another report titled A Statistical methodology for Teacher Evaluation Using Student Achievement Data, Sanders and McLean (undated) conclude that "Henderson's Mixed Model Solution (MMS) procedure which provides BLUP of the random effects while providing opportunity for the inclusion of both continuous and classification fixed effects is a sufficient procedure to provide the flexibility necessary to handle the diversity of specific models which will be encountered in teacher evaluation." This is language designed to communicate to professional statisticians and researchers, not the average educator, school board member, or legislator.

In other documents designed to describe the value-added system, the language contains only generalities without any sense of how the calculations are determined. See, for example, "An Overview of the Tennessee Value-Added Assessment System" by Sanders and Horn (no date), which includes the following one sentence explanation: "TVAAS is a statistical process which provides measures of the influence that school systems, schools, and teachers have on indicators of student learning." And, later, in the text:

> Rigorous statistical theory underpins the TVAAS model. Sanders and McLean demonstrated that Henderson's mixed-model methodology ... when applied in the context of educational outcome assessment, would eliminate most of the statistical problems previously identified as impediments to the use of student achievement data as part of an assessment process.

In other words, from the documents I reviewed, I found that the earliest documents (and the most frequent references) were written in statistical language. The more recent (and less numerous) documents were written with general explanations that leave the reader without an understanding of how the calculations are performed. The problem this creates is that those people who are most affected by the statistical calculations do not have an understanding of what is happening. This breeds suspicion and lack of support.

There are two ways to address this problem. First, if the sophisticated statistical procedures using the so-called "mixed model" are to be continued, efforts need to be made to guarantee to all audiences that the procedures are well-documented, do what they seek to do, and are accurately applied. The review of the value-added procedure by the three out-of-state consultants which led to this paper is a step in the right direction. However, I think it would be useful to do three more things. First, an audit procedure needs to be established to guarantee that all calculations are done correctly. Second, a technical advisory committee should be established to monitor the implementation of the

statistical analyses and to advise the Commissioner on needed changes to the model or its implementation. Third, greater efforts must be made to disseminate information about the value-added system in language understandable to the average educator and citizen.

Evaluation of Teachers. One controversial feature of the Education Improvement Act is a provision that so-called "teacher effects" shall be calculated annually through the value-added system. These calculations are to be provided for teachers in grades three through eight based on the results of the TCAP tests in the five subject areas that are measured. The statute specifies that the estimations of teacher effects will not be a public record but will be made available to "the specific teacher, the teacher's appropriate administrators as designated by the local Board of Education, and school board members" (Section (g)(6)). Furthermore, the legislation presupposes that the teacher effects model can be extended to all high school academic subject areas.

The value-added model is asserted by Sanders as being capable of being properly applied at the classroom teacher level, and, indeed, certain elements of the testing program are specifically designed to gather information with which to support these analyses. Apparently, the value-added model has never been field-tested statewide at the classroom teacher level, so it is not known exactly how and if the model will work with this application.

Certainly, legislators, educational administrators, and parents want to know that classroom teachers are being rigorously evaluated with incapable teachers being terminated. No one wants a child to be assigned to a teacher who is underprepared or ineffective in the classroom. The EIA anticipates that the value-added system is capable of providing an objective and analytic approach to the task of evaluating teachers. In Section 59 of the EIA, the Legislature gave local boards of education the responsibility of establishing guidelines and criteria to evaluate certificated personnel including (1) classroom observations, (2) review of prior evaluations, (3) personal conferences, and (4) other criteria including the Sanders (i.e., the value-added) model. In enlightened school systems, this framework can be implemented in a manner that provides a broad basis for evaluating a teacher's performance. In other systems, this may not happen, and the teacher will be evaluated with too much emphasis on the value-added index.

Whether the value-added indices should be used in teacher evaluation turns on several statistical and psychometric questions that have been analyzed throughout this paper. These discussions will not be repeated here. Alternatively, I will discuss several issues related to teacher evaluation that have not been addressed.

The question of what constitutes a good evaluation of a teacher must be considered. The EIA does not specify the characteristics of a good evaluation -- this is left to the local boards of education. In reviewing the professional literature, several important points emerge. Peterson and Kauchak (1982) make a distinction between evaluation (determining the merit of a person's performance) and research (determining generalizable truths for decision-makers). They describe how teacher evaluations can be

formative (feedback for improvement) or summative (judgments for employment decisions). Teacher evaluations are seen as quite complex affairs dependent upon decisions concerning which data to collect and quantify. The authors caution that "It is rare, in something as complicated as teaching, that a numerical decision by itself is adequate for judgment" (p. 8). Multiple data sources are seen as necessary.

Larson, writing for the NASSP Bulletin, suggests that we evaluate teachers because evaluation has been shown to improve student learning (1984). He stresses that any evaluation strategy should "be designed to protect a teacher from unjust criticism as well as to provide specific information to the teacher whose work is unsatisfactory" (p. 16).

Ramos-Cancel and Duttweiler (1986) call our attention to the teacher's right to due process in any teacher evaluation system. Evaluations should be fair and unbiased. The methodology should be capable of differentiating between superior performance and that of lesser levels. The systems must be credible to and accepted by those being evaluated.

The use of student achievement data as part of any teacher evaluation system is controversial. Haertel (1986) suggested that use of student achievement data in evaluating teachers holds promise but is fraught with risks. Berk (1988) wrote an article for the Journal of Personnel Evaluation in Education which was titled, "Fifty Reasons Why Student Achievement Gain Does Not Mean Teacher Effectiveness." He concluded that there are approximately 50 factors affecting students' achievement independent of what the teacher does:

1. There are more than 40 student, school, and test characteristics that cannot be controlled by the teacher.
2. There are 11 sources of invalidity of the pretest-posttest design that can increase achievement gain.
3. The net effect of all of these factors is to produce a sizable gain in achievement which cannot be attributed to the teacher or to classroom instruction.

Darling-Hammond (1984) was emphatic in stating that using student achievement measures to evaluate teachers meant "substituting very limited measures...for professional judgment while sacrificing some of the most important goals of education for anemic proxies that are cheap, easy, and bureaucratically convenient."

Redfield, Craig, and Elliott (1989) studied evaluation systems in Kentucky and Georgia and concluded that test scores, in isolation, "cannot be used to defensibly evaluate teachers..." They listed six reasons for reaching this conclusion:

1. Standardized tests are designed to measure student performance, not teachers'.

2. Tests are not available to measure content taught by all teachers.
3. There are valued outcomes that cannot be measured by standardized tests.
4. Use of average gain is not a reasonable expectation for non-average students.
5. Many factors affecting student achievement cannot be controlled by teachers.
6. Not all of a student's learning can be attributed to a single teacher.

In another study, Redfield (1987) interviewed parents to see what they thought about use of student achievement data to evaluate teachers. The consensus was that teachers should not be evaluated solely on the basis of student achievement data, but such data should be considered.

Any teacher evaluation system in Tennessee should not ignore the basic question of whether students are learning anything while under the influence of a given teacher. However, teaching is an art form and does not follow an industrial model in which all inputs, processes, and outputs can be quantified in a reliable manner. I suggest that the building principal is responsible for evaluating his/her staff. The school system should mandate evaluation systems that are designed to consider many aspects of a teacher's performance. Teachers will not all generate the same improvements in achievement for many reasons. Some teachers have better control of the content. Some are better organized in the classroom. Some are better speakers. Some work better with students and have more loving, supportive personalities.

The value-added methodology is touted as being sufficiently sophisticated to effectively control for all other factors in a student's life. This is presumably derived from the use of scores over time in which a student's previous performance incorporates information about all of the other factors influencing his/her achievement. I am not convinced that this is true. Our measurement devices and procedures are not totally accurate. Teachers often are teaching out of their area of training. Not all teachers can be measured by the present testing program. The proportion of influence of any given teacher in any particular classroom cannot be determined simply by counting hours of teacher-student contact. Fifteen minutes with a superior teacher may clarify concepts for a student that were left confused after ten hours of instruction with a marginal teacher. And, a science teacher with no direct responsibility to "teach math" may, in fact, have considerable influence over a student's mathematics proficiency. The same could be said for an art teacher who improves students' reading ability. The value-added system cannot make determinations of which teacher contributed how much to students' skill.

Student achievement information can and should be considered by a building principal as a teacher is being evaluated. I do not support use of the value-added system for this purpose. I do not support giving the teacher-level value-added information to the school superintendent and school board members because of the potential for misuse and denial of due process rights to the individual teachers. Tennessee's teacher evaluation system should hold the school principal responsible for conducting a well-rounded

evaluation of his/her classroom teachers. If incompetent teachers are retained in a school, I would conclude that the evaluation procedures are incomplete or ineffective or the school principal lacks the skill to successfully and fairly evaluate the teachers.

Last, one must remember that the question of evaluation of teachers is not a matter simply of educational research and statistical methodology. It involves an individual's protected interests in employment. These are rights that cannot be challenged without due process. There is case law available for review that will shed light on what the courts have held in similar circumstances. It would be useful to consider these cases and determine whether the value-added system or one similar to it can withstand legal challenge. Ours is a litigious society, and I suspect that teachers will consider legal action if they believe the evaluation system is irrational or arbitrary.

Recommendations

The value-added system is a key component of the Tennessee accountability program. Apparently, it is not widely understood and is controversial. The controversy can generally be attributed to (1) general opposition to educational accountability, (2) complexity of the statistical approaches, (3) lack of understanding of the value-added system, and (4) concern about use of the value-added system at the teacher level. My recommendations regarding the value-added system are as follows:

1. As the situation now stands, the value-added system seems to drive the entire assessment program. Tennesseans need to decide if the primary purpose of the educational assessment program is going to be the production of value-added numbers with all of the restrictions that imposes. Educational accountability is being achieved in other states with systems that are much less restrictive and are less complicated.

2 If the value-added system is to be continued, it must be translated into language that can be understood better by the users. However, it must be understood that not all aspects of the analysis program are going to be understood. Classroom teachers and administrators probably do not understand how tests are equated or how norms tables are created, but they have learned to use them anyway.

3. I do not recommend use of the value-added system at the classroom teacher level. I recommend that Tennessee strengthen its teacher evaluation processes and hold the building principal responsible for conducting such evaluations. If the school is not making progress, it will be revealed in the accountability results.

4. I recommend that the Department take better control of the overall management and coordination of its accountability and assessment programs to make certain they serve their stated purposes. The

Department should assume responsibility for speaking on behalf of the assessment and accountability results. As a general rule, contractors should not speak for the agency on various policies and programs.

5. Value-added scores at the school and system level should not be reported without including status information. Legislators, citizens, and parents need to know a school's current level of performance as well as its gain scores.

6. Procedures should be established whereby the value-added calculations are audited and verified prior to their release. This may require contracting with an independent contractor to repeat the analyses, or it may be sufficient to establish on-site verification procedures.

VI. Conclusions

Tennessee has taken progressively bold steps over the last decade to improve education and to create student assessment and educational accountability programs. The Educational Improvement Act was a singularly impressive attempt to move Tennessee education ahead. It particularly included provisions to hold systems, schools, and teachers accountable. In reviewing the EIA and its associated programs, I was impressed and encouraged.

Educational accountability can be attained in various ways. Each state will have its own approach that is driven by its history, its traditions, and its educational philosophy. There is no one best way to hold education accountable nor is there only one way to assess students. Tennessee has to assemble its programs in the best way it can. In this light, then, my recommendations are offered for consideration. Each idea has to be weighed and then either adopted, modified, or discarded. The process of considering these ideas is important in that it will either improve Tennessee's programs or strengthen current resolve that the present programs are on the right track.

Tennessee's educational accountability program can be enhanced if a long-range, comprehensive assessment plan is created and implemented. The plan should clearly state an overall statement of assessment philosophy, the primary purpose of each test, the data to be produced, and the uses of the data. In preparing the plan, it would do well to remember that student achievement difficulties do not merely exist in grades K-12. The cost of remediation for underprepared students is much higher if they enter the postsecondary levels. For this reason, I recommend that Tennessee create and implement a comprehensive accountability and assessment program that includes the postsecondary level.

The Legislature should conduct an overall management review of its accountability and assessment programs to determine answers to these questions:

a. Are the various components being adequately coordinated with the existing managerial structure?

b. Are adequate personnel available to accomplish the functions being required?

c. Are the state's bidding and contracting processes adequate to support the ever-changing, fast moving demands of an educational accountability program?

d. Is educational accountability as a concept being applied at the postsecondary level with the same vigor as it is being applied at the K-12 level?

I complete this report with what may be an obvious point, but it is well to remember it. The State of Tennessee is responsible for creating and maintaining an educational system to protect the public welfare. The State has the obligation to establish educational standards for students and to expect the students to achieve the standards. Educators act as an agent of the State in performing their duties to educate the students. The public welfare is protected to the degree that students are graduated who can pursue a meaningful life. Tennessee must have an educational accountability program as a means of monitoring the educational system's effectiveness in meeting this goal.

Selected References

21st century schools program report card. (1995, October). Nashville, Tennessee: Office of the Governor.

Baker, A.P., Xu, D., & Detch, E. (1995, April). The measure of education: A review of the Tennessee Value-Added Assessment System. Nashville, Tennessee: Comptroller of the Treasury, Office of Education Accountability.

Benavides, L. May 31, 1995. State achievement testing faces more tough scrutiny. Nashville Tennessean, no page identified.

Benavides, L. 1995, October 8. Value-added controversy still unresolved. Nashville Tennessean, p. 1A.

Berk, R.A. (1988). Fifty reasons why student achievement gain does not mean teacher effectiveness. Journal of Personnel Evaluation in Education, 1, 345-363.

Brookover, W.B., Schweitzer, J.H., Schneider, J.M., Beady, C.H., Flood, P.K., & Wisenbaker, J.M. (1978). Elementary school social climate and school achievement. American Educational Research Journal, 15, 301-318.

Cannon, A. (no date). Legislators hold line on retaining ed accountability. Murfreesboro, Tennessee, Daily News Journal. p. 1A.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Modd, A.M., Weinfeld, F.D., & York, R.L. (1966). Equality of educational opportunity. Washington, DC: US Government Printing Office.

Contract Between the State of Tennessee and Thomas H. Fisher. (1995, November 3). unpublished, paragraph A.4.

CTB/McGraw-Hill. Technical report, Comprehensive Tests of Basic Skills, Fourth Edition. (1991). Monterey, California: author.

Darling-Hammond, L. (1984, fall). Taking the measure of excellence: The case against basing teaching evaluation on student test scores. American Educator, 26-46.

Debra P. v. Turlington, 474 F. Supp. 244 (MD Fla. 1979).

Educational Improvement Act, Chapter No. 535 enacted March 11, 1992, by the Tennessee Legislature.

Green, P.J., Dugoni, B. L., Ingels, S.J., & Quinn, P. (1995, June). Trends among high school seniors, 1972-1992. Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Haertel, E. (1986, spring). The valid use of student performance measures for teacher evaluation. Educational Evaluation and Policy Analysis, 8, 45-60.

Hays, W. L. (1973). Statistics (2nd ed.). New York: Holt, Rinehart and Winston.

Instructions for completing the faculty/student data form. (undated) Nashville, Tennessee: Department of Education. A training videotape designed for classroom teachers.

Larson, R. (1984, February). Teacher performance evaluation -- what are the key elements? NASSP Bulletin, 13-18.

Mayshark, J.F. (Oct. 15, 1995). Value-added testing gives Knox schools mark to shoot for despite mixed scores. Knoxville, Tennessee, News-Sentinel, no page identified.

Mullis, I. V.; Dossey, J. A.; Owen, E. H.; & Phillips, G. W. (April 1993). NAEP 1992 Mathematics Report Card for the Nation and the States. Washington, D.C.: U. S. Department of Education, National Center for Education Statistics.

Mullis, I. V.; Campbell, J. R. & Farstrup, A. E. (September 1993). NAEP 1992 Reading Report Card for the Nation and the States. Washington, D.C.: U. S. Department of Education, National Center for Education Statistics.

Peterson, K. & Kauchak, D. (January, 1982). Teacher evaluation: Perspectives, practices, and promises. (ERIC Document Reproduction Services No. ED 233 996).

Purkey, S.C. and Smith, M.S. (March 1983). Effective schools: a review. Elementary School Journal, 83, 427-52.

Ramos-Cancel , M. L. & Duttweiler, P.C. (November, 1986). Issues in perspective: assessment issues papers. (ERIC Document Reproduction Services No. ED 340 109).

Redfield, D.L. (1987). A comparison of the perspectives of teachers, students, parents, and principals concerning the influences of teaching on students and the use of student outcomes to evaluate teaching. (ERIC Document Reproduction Services No. ED 290 765).

Redfield, D.L., Craig, J.R., & Elliott, J. (1989). The role of student outcomes in dual purpose teacher evaluation systems: A model for meeting top down and bottom up needs. (ERIC Document Reproduction Services No. ED 336 423).

State Board of Education. Rules, regulations and minimum standards for the governance of Tennessee public schools in the state of Tennessee. (1994). Nashville, Tennessee: author.

Sanders, W.L. & McLean, R. A. (undated). A statistical methodology for teacher evaluation using student achievement data. (publisher not identified).

Sanders, W.L. & Horn, S. (1993). An overview of the Tennessee Value-Added Assessment System. Knoxville, Tennessee: University of Tennessee.

Sanders, W.L. & Horn, S. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed-model methodology in educational assessment. Journal of Personnel Evaluation in Education, 8, 299-311.

Schneider, J.F. (undated). An educational-assessment database for the state of Tennessee. (Source unknown, available from the author at the University of Tennessee).

State Testing and Evaluation Center. (no date). Tennessee student test results, 1994-1995. Knoxville, Tennessee: author.

Stroup, W.W. (undated). Assessment of the statistical methodology used in the Tennessee Value-Added Assessment System (TVAAS). Unpublished manuscript.

TEA News. (January 1995). Nashville, Tennessee: Tennessee Education Association.

State Testing and Evaluation Center. Tennessee Comprehensive Assessment Program: Building Level Testing Coordinator's Manual, Form F. (1994). Nashville, Tennessee: author.

State Testing and Evaluation Center. Tennessee Comprehensive Assessment Program: Domains and Item Specifications. (1988). Nashville, Tennessee: author.

State Testing and Evaluation Center. Tennessee Comprehensive Assessment Program: Guide to Test Interpretation. (1993). Nashville, Tennessee: author.

State Testing and Evaluation Center. Tennessee Comprehensive Assessment Program Norm-Referenced Component (CTBS/4). (undated). Nashville, Tennessee: author.

State Testing and Evaluation Center. Tennessee Comprehensive Assessment Program: Student Test Results, 1994-1995. (undated). Nashville, Tennessee: author.

State Testing and Evaluation Center. Tennessee Comprehensive Assessment Program: System Testing Coordinator's Manual, Form F. (1994). Nashville, Tennessee: author.

Test scores one of only ways to judge. (no date available). Murfreesboro, Tennessee, Daily News Journal, no page identified.

Watts, W. (1995, no date available). Cannon students do well on Tennessee report card. Murfreesboro, Tennessee, Daily News Journal, no page.

Watts, W. (1995, no month available). Value-added bad way to evaluate: teachers. Murfreesboro, Tennessee, Daily News Journal, no page identified.

Appendix A
Illustrative Achievement Data by Ethnic Group

System Apple, School One
Grade 4 White

Mathematics Total

Value	Freq.	Percent	Percent	Percent
649	1	4.30	4.30	4.30
664	1	4.30	4.30	8.70
690	1	4.30	4.30	13.00
691	1	4.30	4.30	17.40
700	1	4.30	4.30	21.70
703	1	4.30	4.30	26.10
712	1	4.30	4.30	30.40
734	2	8.70	8.70	39.10
737	3	13.00	13.00	52.20
738	2	8.70	8.70	60.90
747	2	8.70	8.70	69.60
749	1	4.30	4.30	73.90
754	1	4.30	4.30	78.30
756	1	4.30	4.30	82.60
763	1	4.30	4.30	87.00
766	1	4.30	4.30	91.30
777	1	4.30	4.30	95.70
781	1	4.30	4.30	100.00
Total	23	100.00	100.00	

Valid cases 23 Missing cases 0

Appendix A (Continued)

System Apple, School One
Grade 4 Black

Mathematics Total

Value	Freq	Percent	Percent	Percent
530	1	2.6	2.6	2.6
532	1	2.6	2.6	5.3
568	1	2.6	2.6	7.9
584	1	2.6	2.6	10.5
605	1	2.6	2.6	13.2
641	1	2.6	2.6	15.8
643	2	5.1	5.3	21.1
646	1	2.6	2.6	23.7
648	1	2.6	2.6	26.3
661	1	2.6	2.6	28.9
666	1	2.6	2.6	31.6
671	1	2.6	2.6	34.2
675	1	2.6	2.6	36.8
676	1	2.6	2.6	39.5
677	1	2.6	2.6	42.1
681	1	2.6	2.6	44.7
685	2	5.1	5.3	50
687	1	2.6	2.6	52.6
691	2	5.1	5.3	57.9
697	1	2.6	2.6	60.5
700	1	2.6	2.6	63.2
706	1	2.6	2.6	65.8
709	1	2.6	2.6	68.4
712	1	2.6	2.6	71.1
713	1	2.6	2.6	73.7
714	1	2.6	2.6	76.3
716	1	2.6	2.6	78.9
723	1	2.6	2.6	81.6
724	2	5.1	5.3	86.8
736	1	2.6	2.6	89.5
738	1	2.6	2.6	92.1
740	1	2.6	2.6	94.7
753	1	2.6	2.6	97.4
758	1	2.6	2.6	100
	1	2.6		Missing
Total	39	100	100	

Valid cases 38 Missing cases 1

Appendix A (Continued)

System Apple, School One
Grade 4 Other

Mathematics Total

Value	Freq	Percent	Percent	Percent
564	1	4.30	4.30	4.30
575	1	4.30	4.30	8.70
583	1	4.30	4.30	13.00
616	2	8.70	8.70	21.70
629	1	4.30	4.30	26.10
640	1	4.30	4.30	30.40
643	1	4.30	4.30	34.80
648	1	4.30	4.30	39.10
651	1	4.30	4.30	43.50
689	1	4.30	4.30	47.80
700	1	4.30	4.30	52.20
708	2	8.70	8.70	60.90
718	1	4.30	4.30	65.20
720	1	4.30	4.30	69.60
721	1	4.30	4.30	73.90
729	1	4.30	4.30	78.30
738	1	4.30	4.30	82.60
739	1	4.30	4.30	87.00
741	1	4.30	4.30	91.30
765	1	4.30	4.30	95.70
767	1	4.30	4.30	100.00
Total	23	100.00	100.00	

Valid cases 23 Missing cases 0

Appendix A (Continued)

System Orange, School One
Grade 8, White

Mathematics Total

Value	Freq	Percent	Percent	Percent
693	1	1.40	1.40	1.40
696	1	1.40	1.40	2.90
729	1	1.40	1.40	4.30
745	1	1.40	1.40	5.80
747	1	1.40	1.40	7.20
748	1	1.40	1.40	8.70
749	1	1.40	1.40	10.10
750	1	1.40	1.40	11.60
751	1	1.40	1.40	13.00
754	1	1.40	1.40	14.50
757	2	2.90	2.90	17.40
760	1	1.40	1.40	18.80
763	2	2.90	2.90	21.70
769	1	1.40	1.40	23.20
770	1	1.40	1.40	24.60
771	2	2.90	2.90	27.50
773	1	1.40	1.40	29.00
774	1	1.40	1.40	30.40
776	3	4.30	4.30	34.80
777	4	5.80	5.80	40.60
778	1	1.40	1.40	42.00
779	1	1.40	1.40	43.50
781	1	1.40	1.40	44.90
782	1	1.40	1.40	46.40
784	1	1.40	1.40	47.80
789	2	2.90	2.90	50.70
793	1	1.40	1.40	52.20
795	1	1.40	1.40	53.60
797	1	1.40	1.40	55.10
799	1	1.40	1.40	56.50
801	1	1.40	1.40	58.00
804	1	1.40	1.40	59.40
809	2	2.90	2.90	62.30
810	1	1.40	1.40	63.80
813	2	2.90	2.90	66.70
814	2	2.90	2.90	69.60
816	1	1.40	1.40	71.00
821	1	1.40	1.40	72.50
822	1	1.40	1.40	73.90
823	1	1.40	1.40	75.40
828	1	1.40	1.40	76.80
829	1	1.40	1.40	78.30
832	1	1.40	1.40	79.70
835	1	1.40	1.40	81.20
842	1	1.40	1.40	82.60
845	1	1.40	1.40	84.10
850	1	1.40	1.40	85.50
859	1	1.40	1.40	87.00
862	1	1.40	1.40	88.40
865	1	1.40	1.40	89.90
868	1	1.40	1.40	91.30
875	1	1.40	1.40	92.80
879	3	4.30	4.30	97.10
887	1	1.40	1.40	98.60
888	1	1.40	1.40	100.00
Total	69	100.00	100.00	

Valid cases 69 Missing cases 0

Appendic A (Continued)

System Orange, School 10

Grade 8 Black
Mathematics Total

Value	Freq	Percent	Percent	Percent
603	1	1.00	1.00	1.00
626	1	1.00	1.00	2.10
655	1	1.00	1.00	3.10
675	1	1.00	1.00	4.20
677	1	1.00	1.00	5.20
687	1	1.00	1.00	6.30
706	1	1.00	1.00	7.30
712	1	1.00	1.00	8.30
718	1	1.00	1.00	9.40
719	1	1.00	1.00	10.40
727	1	1.00	1.00	11.50
728	1	1.00	1.00	12.50
730	1	1.00	1.00	13.50
732	1	1.00	1.00	14.60
734	1	1.00	1.00	15.60
735	1	1.00	1.00	16.70
736	2	2.10	2.10	18.80
737	1	1.00	1.00	19.80
743	2	2.10	2.10	21.90
744	1	1.00	1.00	22.90
746	1	1.00	1.00	24.00
747	1	1.00	1.00	25.00
750	1	1.00	1.00	26.00
754	2	2.10	2.10	28.10
755	1	1.00	1.00	29.20
756	2	2.10	2.10	31.30
757	1	1.00	1.00	32.30
758	1	1.00	1.00	33.30
759	1	1.00	1.00	34.40
761	2	2.10	2.10	36.50
764	2	2.10	2.10	38.50
765	2	2.10	2.10	40.60
767	3	3.10	3.10	43.80
769	4	4.20	4.20	47.90
771	3	3.10	3.10	51.00
772	3	3.10	3.10	54.20
774	3	3.10	3.10	57.30
775	1	1.00	1.00	58.30
777	3	3.10	3.10	61.50
778	2	2.10	2.10	63.50
779	1	1.00	1.00	64.60
780	1	1.00	1.00	65.60
781	2	2.10	2.10	67.70
782	1	1.00	1.00	68.80
784	1	1.00	1.00	69.80
785	1	1.00	1.00	70.80
786	1	1.00	1.00	71.90
789	1	1.00	1.00	72.90
790	3	3.10	3.10	76.00
792	1	1.00	1.00	77.10
793	1	1.00	1.00	78.10
794	1	1.00	1.00	79.20
798	1	1.00	1.00	80.20
799	1	1.00	1.00	81.30
801	1	1.00	1.00	82.30
803	1	1.00	1.00	83.30
808	1	1.00	1.00	84.40
811	2	2.10	2.10	86.50
818	1	1.00	1.00	87.50
819	2	2.10	2.10	89.60
821	1	1.00	1.00	90.60
822	1	1.00	1.00	91.70
825	1	1.00	1.00	92.70
826	2	2.10	2.10	94.80
827	1	1.00	1.00	95.80
831	1	1.00	1.00	96.90
845	1	1.00	1.00	97.90
853	1	1.00	1.00	99.00
869	1	1.00	1.00	100.00
Total	96	100.00	100.00	

Valid cases 96 Missing cases 0