The Performance Levels and Associated Cut Scores *on the* Pennsylvania System of School Assessment Mathematics and Reading Tests: *A Critical Analysis*

Harris L. Zwerling, J.D., Ph.D.
Assistant Director of Research
Pennsylvania State Education Association
P.O. Box 1724
400 N. Third Street
Harrisburg, PA 17105-1724

Acknowledgements

The author gratefully acknowledges the assistance of Stinson Stroup, Executive Director of PASA, Marcia Bender, Joe Bugden, and Frank Rosenhoover for helping secure participants for the PSSA-commercial standardized test performance comparison study. The participation of the four school district superintendents and the data compilation by their staff also is gratefully, albeit anonymously noted.

Donna Bowers, Tracy Gardner and Elaine Easton of PSEA Research assisted in preparation of tables and typing of sections of this paper. Thanks also to David Helfman, Jon Schmehl and Dave Wazeter of PSEA Research for comments on a previous draft. However, all errors remain the responsibility of the author.

Executive Summary

- "High-stakes" tests have generated a great deal of controversy. However, the broad goal of educational accountability enjoys widespread popular support.
- Few among the public, news media, or elected and non-elected government officials have shown much inclination to delve into the technical details of what is one of the most dramatic exercises of educational policymaking undertaken in the last quarter century: establishing performance category cut scores. Instead, a small group of expert testing advisers and government education officials make these decisions without comprehensive independent scrutiny or much public disclosure.
- Cut scores are what define the "bottom line." They define success or failure. For students, they have the potential to impact class placements, course selections, college decisions and lifetime earnings. For educators, they can affect job benefits and security, as well as curriculum and instructional decisions. For schools and districts, they can result in privatization. For taxpayers, they can even result in upward—or downward—pressure on housing values.
- The Pennsylvania Department of Education used two different methods for establishing cut scores: the bookmark and the borderline groups methods. The bookmark standard setting studies were conducted by PDE and facilitated by CTB/McGraw-Hill.
- Instead of selecting the more fully-documented and carefully executed (bookmark) method for establishing cut scores, the Secretary of Education averaged the cut scores established using these two methods, then increased the cut scores by "a quarter standard error." This data manipulation resulted in tens of thousands of students falling into lower performance categories.
- Over all six tests (two subjects, three grade levels), when compared to the bookmark study cut scores, the cut scores approved by the Secretary **decrease the number of students achieving proficient or better by over 25,000.** The PDE cut scores **increase the number of students failing by more than 42,000.**
- The use of multiple hurdles for determining which students receive diploma seals increases the likelihood that a student with ability above the proficient level in all test subjects will not receive a graduation seal due to measurement error (a "false negative").
- The PSEA-PASA study of the relationship between PSSA performance levels and performance on commercial standardized tests indicates:
    - While there is a strong relationship between performance on the PSSA tests and commercial tests, many students judged to be failing under PSSAs do quite well on other commercial tests.
    - **Below-basic PSSA performers averaged slightly higher SAT performance than those classified as basic,** providing additional evidence that the failure cut score on the 11th grade PSSA Reading test bears re-examination.

The Performance Levels and Associated Cut Scores on the Pennsylvania System of School Assessment Mathematics and Reading Tests: *A Critical Analysis*

Background

Across America the executive and legislative branches of state governments have seized upon student testing as a primary means for achieving educational accountability and reform.

This movement recently reached its zenith with Congress' recent passage of the "No Child Left Behind Act of 2001," reauthorizing ESEA, the Elementary and Secondary Education Act. That legislation signed into law by President Bush on January 8, 2002, requires states to annually test all public school students in reading and math from grades three through eight.

While "high-stakes" tests have generated a great deal of controversy nationwide, the broad goal of educational accountability continues to enjoy widespread popular support. Unfortunately, few among the general public, news media, or elected and non-elected government officials have shown much inclination to delve into the technical details of what is one of the most dramatic exercises of educational policymaking undertaken in the last quarter century. Instead, they have left those matters in the hands of a relatively small group of highly skilled testing experts acting as advisers to government education officials. This paper will take a closer look at recent changes in Pennsylvania's state testing program.

State Board of Education Adopts Regulations Calling for Student Assessments; Proficiency Levels

In January 1999, the Pennsylvania State Board of Education (State Board) published regulations under Chapter 4 intended to "establish rigorous academic standards and assessments to facilitate the improvement of student achievement and to provide parents and communities a measure by which school performance can be determined (Pennsylvania Bulletin, 2001)."

The regulations established the following "levels of proficiency": advanced, proficient, basic, and below basic. The State Board directed the Pennsylvania Department of Education (PDE) to "develop and recommend to the Board for its approval specific criteria for each performance level (Ibid.)." On May 10, 2001, the State Board approved new performance level cut scores and descriptors for the Pennsylvania System of School Assessment (PSSA) Math and Reading tests (see Figure 1).

According to a document posted the following day on the Pennsylvania Department of Education (PDE) website:

> "For the first time next fall, the results of our reading and math assessments will be presented with more than just a test score," (Education Secretary) Zogby said when he introduced the levels. "We'll now have performance levels -- clear descriptions of our students' performance that will help parents, teachers and students better understand what those scores really mean" (Pennsylvania Department of Education, 2001a).

The PDE document continued by noting, "The four levels are: advanced (superior academic performance); proficient (satisfactory academic performance); basic (marginal academic performance); and below basic (inadequate academic performance)(Ibid.)."

Figure 1

PSSA Levels of Proficiency

Advanced……. **Superior academic performance** indicating an in-depth understanding and exemplary display of the skills included in Pennsylvania's Academic Standards.

Proficient…….. **Satisfactory academic performance** indicating a solid understanding and adequate display of the skills included in Pennsylvania's Academic Standards.

Basic……. **Marginal academic performance, work approaching, but not yet reaching, satisfactory performance**. Performance indicates a partial understanding and limited display of the skills included in the Pennsylvania's Academic Standards, and the student may need additional instructional opportunities and/or increased student academic commitment to achieve the Proficient Level.

Below Basic…….. **Inadequate academic performance** that indicates little understanding and minimal display of the skills included in the Pennsylvania Academic Content Standards. There is a major need for additional instructional opportunities and/or increased student academic commitment to achieve the Proficient Level. (Pennsylvania Bulletin, 2001)

Performance on 11th Grade Assessments Linked to Diploma Seals

While the performance levels and associated cut scores were intended to aid in interpretation of test results at all grade levels, they will hold greater significance for high school students.

> Students graduating in 2003 who score at proficient or advanced levels on the 11th grade state assessments will be eligible to receive special seals on their diplomas: Seals of Proficiency and Seals of Distinction, respectively. The seals will signify that students have achieved high levels of excellence on the state's reading, writing and math assessments. Graduation requirements still will be established by local school districts (Pennsylvania Department of Education, 2001a).

In their May 9, 2001 presentation to the State Board, PDE staffers anticipated that the cut scores for each of these levels initially would roughly correspond with the quartile distributions used to report previous years' scores. Given historical patterns of small upward annual movement in the PSSA scores, one would expect that nearly half of the state's students will continue to score at basic or below basic.

This last autumn's release of the results of the 2000-01 PSSA tests indicated that across the six Math and Reading tests a range of 39.9 to 52.1 percent of the state's students scored at the basic or below basic levels (Pennsylvania Department of Education, 2001b).

After the adoption of the performance levels, press reports have employed these new descriptors when discussing the performance of public schools throughout the Commonwealth. However, to our knowledge, none in the mass media have examined the basis for the claim that the performance levels "will help parents, teachers and students better understand what those scores really mean" (Pennsylvania Department of Education, 2001a).

Criterion-referenced v. Norm-referenced Assessments

One purpose of the adoption of these new performance descriptions was to move away from a norm-referenced reporting system, i.e., one in which students and school results were described in terms of scaled scores and quartiles, to one which was criterion-referenced. In the latter approach, student and school results could be described with reference to the state's academic standards. Then, one supposedly could meaningfully characterize performance in terms of academic failure or proficiency with respect to the attainment of these standards[1]. In a norm-referenced system, student and

[1] Note the term "failure" does not appear in the definitions listed in Figure 1. However, former Education Secretary Eugene Hickok, current Secretary Zogby, and Governor Schweiker have repeatedly referred to students falling in the bottom quartile of the PSSAs, and since adoption of the performance levels, students

school performance is judged relative to that of other students and schools, respectively. In a criterion-referenced system, it is theoretically possible for all to "pass" or "fail," or achieve at an advanced level.

PSEA Seeks Technical Report; PDE Denies Access

Among other reasons, PSEA was concerned that the regulations' definition of basic performance may lead many to conclude that nearly half the students in the Commonwealth are "not yet reaching...satisfactory performance," a phrase that, read alone, suggests failure. Consequently, in late May of 2001 PSEA initiated a series of contacts with PDE officials intended to clarify these concerns and inquire regarding the technical bases for the establishment of the specific cut scores demarcating the performance levels for each test.

On June 28, 2001, four PSEA staffers led by Executive Director Carolyn Dumaresq, met with Secretary of Education Charles Zogby and the PDE staff responsible for the design and implementation of the new PSSA performance levels.

Education Secretary Zogby assured PSEA representatives that when calculating the Empowerment List under Act 16,[2] the Department would not count students scoring in the "basic" category as 'failing." However, he did indicate that it was the Administration's goal to move all students towards proficiency.

Dr. Leonard Lock of PDE also indicated that the Department was recently in contact with a prominent academic testing expert who proposed her own consequential validation study of the performance levels. During the meeting, PSEA hand delivered a letter detailing its request for technical information relating to the establishment of the performance levels and related cut scores (See Appendix A).

PSEA Files Right to Know Request; PDE Provides Access

Approximately one month after the meeting, PDE denied PSEA access to the requested information citing ongoing litigation that PSEA has maintained regarding the application of Act 16. PSEA initiated a "Right to Know" request and was given access to PDE's documentation on October 12, 2001 and examined the materials on October 23rd.

While awaiting access to the technical documentation, PSEA decided to undertake its own validation study and solicited the assistance of Stinson Stroup, executive director of the Pennsylvania Association of School

---

falling in the below basic category, as "failing" math and reading. (See Commonwealth of Pennsylvania News Release, 2001a, 2001b, 2002; Executive Summary, 2002)

[2] Act 16 of 2000, amending the Act of MARCH 10, 1949 (P.L.30, NO.14).

Administrators (PASA). Ten Pennsylvania school districts were contacted to participate in the study; four agreed to participate.

Overview of this Study

This report is divided into three main sections.

Part 1 discusses the technical information provided by PDE to the public on its website and before the State Board of Education as well as that provided in response to the PSEA Right to Know request. It assesses the adequacy of the cut score-setting process.

Part 2 contains the joint PSEA-PASA study of the relationship between PSSA performance levels and performance on commercial standardized tests. This was done to provide a gauge of the external validity of the cut scores.

Part 3 presents final conclusions and recommendations[3].

I. Setting Cut Scores: Standard Setting Technical Reports and Other Arcane, but Crucial, Matters

General Observations about Setting Cut Scores

Various experts have noted the subjective or constructed nature of cut scores. An outsider delving into this field for the first time will be impressed by both its technical sophistication as well as the candor with which experts acknowledge subjective influences.

> There is general agreement now that cutscores are constructed, not found. There is no "true" cutscore that researchers could find if only they had unlimited funding and time and could run a theoretically perfect study. …
>
> Standards are to be considered acceptable if they follow a due process model consisting of three aspects: a legitimate purpose, adequate notice, and fundamental fairness. (Zieky, 2001, p. 45)

While it is beyond the scope of this effort to examine all aspects of "fundamental fairness," this report will examine the extent to which the PSSA Math and Reading standards setting were well designed and carefully conducted.

First, it will be instructive to examine the PDE's description of the standard setting process, then view a brief chronology of the main steps taken.

[3] This study does not consider questions regarding the appropriateness of using the PSSA tests as a high stakes performance measure for students, schools, or teachers.

Figure 2

STEPS IN THE PERFORMANCE STANDARD SETTING PROCESS

1) General Definition Development. As prescribed by Chapter 4 Regulation, student achievement is to be reported in one of four levels: advanced, proficient, basic, and below basic levels of performance. To accomplish the task of defining each level, a survey of over 1,700 educators as well as business, professional education, and parent associations was completed to provide these general definitions. The survey was conducted in the Spring 2000.

2) Performance Standard Setting. Two statistical procedures were utilized for the Pennsylvania performance standard setting processes: The bookmark and borderline groups methods. The mathematics bookmark meeting was held in October 1999, and the reading bookmark meeting occurred in September 2000. Parallel procedures were followed at both content area bookmark meetings. All districts had an opportunity to participate in this process via a general announcement to all districts.

The **bookmark method** entailed a representative sample, predominantly comprised of Pennsylvania teachers, examining PSSA "ordered item booklets" (parents, educational organizations, business representatives, and higher education also were represented). More specifically, the bookmark procedure involved teachers' examination of PSSA test booklets arranged in order from high p-value (easy) to low p-value (difficult) items. (In (t)he case of reading, the passages were included as well.) Approximately 70 teachers in (e)ach content area made determinations of advanced, proficient, and basic "cut score" academic achievement levels by placing a "bookmark" at a point in the PSSA booklet that represented each of these levels. That is, the bookmark method was premised upon Pennsylvania teachers' identification of what constitutes a student's academic work to be deemed advanced, proficient, and basic as represented by the PSSA knowledge and skills that comprise these academic levels. To assure content validity, the PSSA items used for the bookmark meetings were comprised of a sample of items that measures all the Pennsylvania Academic Mathematics and Reading Standards. Both PSSA open-ended and selected response items were included in this bookmark procedure.

To maximize standard setting accuracy, the **borderline groups method** was also implemented. This method examined teachers' determinations of their students' academic achievement and then related these ratings to the student's actual PSSA score. This method utilized teachers' ratings of each of their student's academic achievement level and then linked this data with the student's actual year 2000 PSSA scaled score. That is, teachers rated their students in one of seven categories: clearly advanced, borderline advanced/proficient, clearly proficient, borderline proficient/basic, clearly basic, borderline basic/below basic, or clearly below basic. Then, the average actual achieved PSSA scaled score, as well as related statistics, was computed for each of the students in these rated categories. For mathematics and reading combined (across grades 5, 8, and 11), teachers classified a total of 12,536 students' academic achievement. The borderline groups teachers were generally members of the Advisory Committees or their district level colleagues.

Both procedures have been used by other states or large scale assessments as well as companies to set academic performance standards.

3) Presentation to the Secretary of Education. In accordance with chapter 4 Regulation, the results of the bookmark and borderline groups methods were presented to the Secretary for examination and concurrence for subsequent presentation to the State Board of Education.

4) Survey of Performance Expectations. As a final confirmatory step before consideration by the State Board, approximately 30 teachers were reconvened in April 2001 to continue writing specific descriptors and to validate the performance level expectations and associated scores. Overall, over 90% of the teacher ratings (totaled across advanced, proficient, basic, and below basic) indicated that the results of the performance standard setting work were at least acceptable.

5) Specific Descriptor Development. A final step in the performance standard setting process is the development of specific descriptors. Currently, preliminary specific descriptors have been developed based upon teacher groups reviewing sample PSSA test question responses that represent advanced, proficient, basic, and below basic work for all the mathematics and reading academic standards. In the upcoming weeks, teacher groups will be reconvened to provide greater detail and refinement regarding these specific descriptors. (These further detailed specific descriptors will be available in July.) In this manner the specific descriptors will be finalized for use by teachers across the Commonwealth. (Pennsylvania Department of Education, 2001c)

PDE Describes Performance Standard Setting Process to State Board

The unedited paragraphs in Figure 2 were included in the written materials PDE provided to the State Board of Education during its public study session. They provide the Department's summary of the steps followed in their standard setting process. (State Board of Education Study Session, May 9, 2001)

PDE Chronology: Establishing Cut Scores

The following chronology highlights the process used by PDE in establishing cut scores for each of the performance levels.

1. **October 1998**: the Pennsylvania State Board of Education adopted regulations (published January 16, 1999) directing the PDE to "develop and recommend to the Board for approval specific criteria for advanced, proficient, basic and below basic levels of performance." [Ch. 4 sec. 4.51 (b) (4)].
2. **October 4-9, 1999**: PSSA Mathematics bookmark standard setting study was conducted by PDE and facilitated by CTB/McGraw-Hill.
3. **Fall 1999**: PSSA borderline groups standard setting studies for Mathematics was conducted by survey in the fall of 1999. PDE identified the participants from among the members of the Mathematics Assessment Advisory Committee. Survey results were compiled and analyzed by CTB/McGraw-Hill.
4. **September 25-29, 2000**: PSSA Reading bookmark standard setting study conducted by PDE and facilitated by CTB/McGraw-Hill.
5. **Fall 2000**: PSSA borderline groups standard setting studies for Reading were conducted by survey. PDE identified the participants from among the members of the Reading Assessment Advisory Committee. Survey results were compiled and analyzed by CTB/McGraw-Hill.
6. **March 21, 2001**: a PDE "Draft" spreadsheet indicates that the Department averaged the cut scores resulting from the bookmark and borderline groups studies. Three average cut scores were reduced 1 standard error of measurement (Reading 11 advanced, Math 8 proficient, and Math 8 basic).
7. **April 16, 2001**: a later spreadsheet also marked "Draft" increased average cut scores "by one-quarter standard error."
8. **April 18, 2001**: adjusted average PSSA cut scores were approved by the Secretary of Education. This accepted the cut scores calculated on April 16 with the three exceptions:
    - **a.** the Reading 11 advanced score was decreased by 40 points,
    - **b.** the Math 8 proficient score was decreased by 30 points, and

c. the Math 8 basic cut score was decreased by 40 points.

(Thus, the Secretary's recommended cut scores represent a hybrid of the March 21 and April 16 drafts.)

9. **May 10, 2001**: the State Board of Education adopted PSSA cut scores and performance levels.
10. **September 25, 2001**: the CTB/McGraw-Hill Reading Standard Setting Technical Report (dated August) and Mathematics Standard Setting Technical Report (dated September) were received by the Department of Education—over four months *after* the cut scores were adopted by the State Board.

PDE Sets General Definition: Feedback Sought from 1,700

In the materials provided to the public during their May 9 presentation to the State Board, PDE staff went to some lengths to describe the consultation process by which the performance level descriptions were adopted.

In fact, their handout at that meeting devoted seventeen pages to that part of the standard-setting process, but only seven pages to the bookmark, borderline and post-standard setting studies.

As noted in their handout, the PDE

- solicited input from 1,700 teachers, superintendents, and community members regarding the written performance level descriptions,
- reported that 850 responded, and
- listed up to 834 respondents in its response category summary table.

Among the respondents

- 328 of the respondents were district superintendents and
- 48 were primary or secondary school teachers (Pennsylvania Department of Education, 2001c).

The response indicated a high level of satisfaction with the descriptors, one that held up across categories of respondents. This appears to be the only instance where the documentation provided by PDE shows anything close to "thousands" of educators and parents actually actively participating in part of the standard-setting process.

Establishing Cut Scores: It's a Subjective—and Mysterious—Process

At the May State Board meeting, Department officials also provided a description of the two procedures used for setting the actual cut scores that would demarcate the boundaries between adjacent performance levels.

It would not be until several months later with the release of the Technical Reports prepared by CTB/McGraw-Hill, the contractor that conducted the standard setting studies, that a clearer picture would emerge of the judgments made in establishing lines between passing and failing, marginal performance and that meriting a seal of proficiency.

Why do the Cut Scores Matter?

To be blunt, the cut scores are what define the "bottom line." They define success or failure. Setting cut scores is an enormously significant step. For students, it has the potential to impact class placements, course selections, college decisions, and lifetime earnings. For educators, it can affect job benefits and security, as well as curriculum and instructional decisions. For schools and districts, it can result in privatization. For taxpayers, it can even result in upward—or downward—pressure on housing values.

The lack of public discussion of these arcane statistical and policymaking matters is understandable, but that should not be taken as a sign of their relative importance to the validity of this accountability system.

Why did PDE and its Testing Consultants Overstate Teacher Involvement?

For some unknown reason, both CTB/McGraw-Hill and the Department overstated the involvement of teachers in the process.

> The Pennsylvania System of School Assessment cutscores were determined by using information gathered from teachers and administrators in two studies conducted in 2000…The participants for each study were identified by the Pennsylvania Department of Education. Approximately 24 participants per grade participated in the grades 5, 8, and 11 Bookmark standard setting study **and several thousand teachers participated in the Borderline Groups study** (which was conducted by survey). (CTB/McGraw-Hill, 2001a, Section A) *(Emphasis added.)*

Immediately following the State Board's May meeting, the PDE posted on their website the following description of the standard setting process.

> The state Department of Education, along with nearly 1,000 educators, parents, and community and business leaders, used two statistical standard-setting procedures **to develop the performance levels** -- the bookmark method and the borderline groups method (Pennsylvania Department of Education, 2001a). *(Emphasis added.)*

Upon closer examination, the cast of thousands (CTB/McGraw-Hill, 2001a) or nearly a thousand (Pennsylvania Department of Education, 2001a) actually played a much more limited role than suggested by the Technical Reports or

PDE. While 850 (including but 48 teachers) responded to a mail survey concerning the proposed performance level **descriptions**, in contrast, the numbers involved in the **developing the performance levels** (that is, conducting the cut score-setting studies) appear to be dramatically smaller.

The Bookmark Method is Determined by Small Groups

Panels using the bookmark procedure to set cut scores were comprised of teachers, administrators, parents, and business and educational organization representatives selected by the PDE. The numbers of panel members setting cut scores for each test were:

- Math, Grade 5 25
- Math, Grade 8 21
- Math, Grade 11 22
- Reading, Grade 5 22
- Reading, Grade 8 21
- Reading, Grade 11 20

This averaged approximately 22 members per panel (CTB/McGraw-Hill, 2001a and 2001b). So an average of 22 people set the cut scores in the bookmark standard setting studies for each grade level test. However, this actually is slightly higher than the recommended number of panelists typically employed in this type of standard setting (Mitzel, et al., 2001, p. 250).

Borderline Groups Cut Scores Based on 12,536 Students—or Not?

In contrast, the apparent overstatement of the borderline groups study participation levels should cause greater concern. In the document presented to the State Board in May 2001, PDE stated regarding the borderline groups study:

> "For mathematics and reading combined (across grades 5, 8, and 11), teachers classified a total of 12,536 students' academic achievement (Pennsylvania Department of Education, 2001c)."

Given this context, the State Board of Education might be surprised to learn that the borderline groups failure cut score for 11th grade reading was established based on the evaluations of two students statewide made by an unreported number of teachers. The failure cut scores for Reading 5 and Reading 8 were based on the evaluations of 147 and 46 students, respectively.

As will be discussed below in greater detail, the use of the borderline groups study substantially increased the overall failure cut scores (and failure rates) for four of the six PSSA Math and Reading tests.

How Many Teachers Provided Ratings?

While the Reading Technical Report states "several thousand" teachers participated in the borderline group surveys (CTB/McGraw-Hill, 2001a, Section A, first page), neither the Mathematics and Reading Technical reports nor the PDE document (Pennsylvania Department of Education, 2001c) provides the actual number of participating teachers providing the ratings.

Reading Survey Response Rates were Low

The omission is odd, given that the teachers were the actual raters. Instead, the numbers of participating schools and students rated are reported. An examination of the tables provided in the Technical Reports indicates that the cast of several thousands in all likelihood was much smaller, particularly when we examine the numbers involved in responding to the survey by grade level and test.

The Technical Reports indicate that Reading surveys were returned from only:

- 14 schools for the 5th grade.
- 8 schools for the 8th grade.
- **2 schools for the entire 11th grade cut score study**.

Again, it would seem that the number of teachers responding would be the most relevant indicator of the representativeness of the sample and the number of independent raters involved.

The Technical Reports describe the response rates for Reading as "low" from 10 to 32 percent returned (CTB/McGraw-Hill, 2001a, p. I-1). However, the tables referenced indicate the number of schools replying, not the numbers of teachers, who, according to the Reports, "completed and returned the survey(s)."

Mathematics Survey Response Rates were Higher

Reported response rates for the Mathematics borderline surveys were much better, survey responses rates by grade level ranged from 48 to 69 percent (CTB/McGraw-Hill, 2001b, p. J-1). In all, the Technical Reports indicate that Mathematics surveys were returned from:

- 50 schools for the 5th grade.
- 12 schools for the 8th grade.
- 8 schools for the 11th grade cut score study.

Technical Reports Lacking Important Demographic Information

One can only speculate why PDE consistently would choose to highlight the largest possible number of standard-setting participants in their public disclosures. Perhaps their underlying concern is the legitimacy conferred by the standard-setting panels. If the public believes the panels are representative

of the education profession, then perceived legitimacy is enhanced. For that reason it is unfortunate that the Technical Reports provide little demographic information concerning the panelists.

The documentation PDE provided at the May State Board meeting does list the names of districts and other educational organizations that provided participants for the bookmark and borderline studies.

The Technical Reports provide little demographic information about the bookmark panelists and none about the actual participants in the borderline groups survey. The latter is of particular concern given the low response rates for the Reading study. The participants in the latter may have come from an unrepresentative segment of the original sample, which itself may have been unrepresentative of professional educators[4].

PDE staff selected the bookmark study participants from among Pennsylvania's school administrators and teachers. The borderline groups participants were volunteers (i.e., self-selected) from among the members of the State's Reading and Math Assessment Advisory Committees.

While it is common standard-setting to employ raters who have a deep familiarity with a state's academic standards, it would be reasonable to expect evidence of how reflective the views of the particular subset of participants are of their respective advisory committees as well as those of the much larger community of Pennsylvania's professional educators.

PDE Fails to Meet its Own Experts' Standards

Dr. Ronald Hambleton, an internationally renowned expert on testing and standard-setting, and not incidentally, one that PDE's staff prominently mentioned as part of their Technical Advisory Oversight Committee, seems clear on this point:

> (8) Were the qualifications and other relevant demographic data about the panelists collected? (This information is needed to fully inform reviewers about the suitability and composition of the panel setting the performance standards. … Even the panelists motivation for participation may be relevant information.) (Hambleton 2001, p.110)

If this information was collected, none was contained in the Technical Documentation prepared by CTB/McGraw-Hill.

Dr. Daniel Lewis, manager of Strategic Research Programs at CTB/McGraw-Hill, who directed the PSSA standard setting studies, has acknowledged in other written work the possibility of bias introduced by the behavior of panelists.

[4] Moreover, the Technical Reports provide relatively little procedural information. For instance, was any attempt made to determine whether borderline panelists had access to actual student PSSA scores?

Mitzel, (Dr. Daniel) Lewis, Patz, and Green (2001, p. 254) note the possibility of strategic behavior by some bookmark standard-setting panelists (i.e., choosing extreme scores in an attempt to influence the group results). Given this, why did the Technical Reports fail to examine this and other possible sources of selection bias in the choice of panelists by the PDE in the bookmark panels and their self-selection into the borderline studies?

What Benefit Is There to Using Both Methods?

It is clear that of the two standard-setting methods used, the PDE devoted considerably more time, effort, and resources to designing, staging, analyzing, and documenting the bookmark standard setting studies. The Department went so far as to contract with CTB/McGraw-Hill and to bring Dr. Daniel Lewis, one of the method's developers, in to facilitate the studies.

One of the purported advantages of the bookmark method is that it can produce valid and reliable performance standards for both the selected response (multiple choice) and constructed response type questions that appear on assessments such as the PSSA tests (Mitzell, Lewis, et al., 2001)[5].

As quoted above, PDE staff reported to the State Board that they undertook a borderline standard setting study to "maximize standard setting accuracy." Michael Kane, of the University of Wisconsin, offered another rationale for using two different standard procedures.

> Another way to check on the appropriateness of the cutscore resulting from a standard-setting study would be to conduct another standard-setting study on the same test using a different method. … If the two approaches agree, we have more confidence in the resulting cutscores than we would if either method were used alone. …
>
> A lack of agreement between two standard-setting studies using different methods should not be very surprising, because different methods ask participants to use different kinds of data (i.e., different item characteristics or student performances in different ways). Nevertheless, if we consider the methods to be exchangeable in the sense that the resulting cutscores are interpreted in the same way, large

[5] While this report does not question the validity or promise of the bookmark technique for setting performance standards, it is instructive to consider the recent comments of expert and PDE Technical Advisory Committee Member Ronald Hambleton:

> Finally, methods for setting performance standards on educational assessments using the multiple-choice item format are well developed and steps for implementation are generally clear (citations omitted). … On the other hand, standard setting methods for educational assessments that include constructed response items such as writing samples and performance tasks are not as well developed at this time, and certainly none have been fully researched and validated (citations omitted) (Hambleton, 2001, p. 93).

discrepancies tend to undermine confidence in both cutscores (Kane, 2001, p. 75).

Similarities—and Meaningful Differences—Between the Two Methods

How much did the two methods agree? The answer depends both on which cut score one examines and what technique one uses to compare the two results.

Technical Reports Cite Agreement; Focus on Proficient Level Students

The first of three ways available to compare results using the two different methods is to compare percentages of test-takers at each performance level.

According to the description in the Reading Technical Report, "results indicate reasonably good agreement between the two methods (CTB/McGraw-Hill 2001a, p. I-2)."

The Mathematics Technical Report finds "good agreement between the two methods" when used to set the math cut scores (CTB/McGraw-Hill 2001b, p. J-2).

Both technical reports then provide a preliminary comparison of "the differences in the results of the two standard setting procedures in terms of the estimated percent of students in each performance level." However, both reports only discuss the relative proportions of students performing at or above the proficient level.

Analysis of PDE Data: Failure Rate Up Substantially with Borderline Method

Using PDE data, PSEA's calculations indicate that the borderline studies result in an average across all six PSSA Math and Reading tests of 1.4 percent fewer students scoring at or above proficient compared with the results from the bookmark method (based on the 2000 PSSA score distributions).

However, the borderline method produces an average failure rate (proportion of students scoring below basic) that is 8.6 percent above that of the bookmark method across all tests. Given an average failure rate of 20.6 percent under the bookmark method, this failure rate of 29.2 percent **represents a 42 percent increase in the number of students who would fail (See Appendix D).**[6]

[6] These calculations were made by averaging across all six tests the percentage of students who would score "below basic" using the borderline cut score and comparing it with the average percentage scoring "below basic" using the bookmark cut score.

Technical Report: Cross-tabs Confirm Similarities and Differences

The Technical Reports used **a second** means to compare the bookmark and borderline results. Each report provided a cross tabulation of the results from the two methods at each test and grade level (CTB/McGraw-Hill, 2001a, p. I-11; 20001b, p. J-11). Those tables are reproduced in Appendices B and C.

These tables take the students rated in the seven categories used in the borderline study (listed across the first row of each table) and compare those classifications with how they would have been classified using the bookmark cut scores and categories (listed down the first column of each table).

The bolded numbers show agreement between the two methods. One number is bolded in columns falling under "Clearly Advanced," "Clearly Proficient," "Clearly Basic," or "Clearly Below Basic." In the columns marking the "Borderline" between two performance level categories, two numbers are bolded one for each of the adjacent performance level categories.

Looking at the Math tables first, as noted in the Technical Reports, one can see that the individual or paired bold numbers are the largest in each column (CTB/McGraw-Hill, 2001b, p. J-2).

However, looking closely we can also see that in the Math 5 Table less than half (48 percent) of the students classified by the borderline study as clearly below basic would have been classified as failing by the bookmark method. **In other words, under the bookmark method, fewer students appear to be failing and many more would be performing at higher levels.**

A review of the Reading tables reveals some meaningful differences between the results of the two methods.

Looking at the 5th Grade Reading tables we can see:

- only 45 percent of those classified as **below basic** in the borderline study would have **failed** under the bookmark method.
- 36 percent classified as **clearly basic** would have been classified as **basic** under the bookmark method, while 57 percent would have been classified as **proficient or advanced** under the bookmark method.

At the 8th Grade Reading level:

- only 55 percent of those classified by the borderline study as **below basic** would have **failed** under the bookmark cut.
- 48 percent classified as **clearly basic** would have been basic under the bookmark method. Another 29 percent would have reached **proficiency** under the bookmark cut (i.e., many scored in higher categories under the bookmark procedure).

The 11th grade Reading table highlights the most questionable results from the borderline study.

- None of the five students classified as **clearly below basic** in the 11th grade borderline Reading study would have **failed under** the bookmark method.
- Thirty-six percent of those classified as **clearly basic** in the 11th grade borderline study would have been classified as **basic** under the bookmark method, while 61 percent classified as **clearly basic** would have attained **proficiency** under the bookmark method (i.e., many scored in higher categories under the bookmark system).

It is clear that for many students the score-setting method selected (borderline or bookmark) does make a difference in how they are labeled based on their PSSA test performance.

Comparing Different Methods by Analyzing Standard Errors

To understand **a third** means of gauging agreement between methods, consider what Dr. Lewis of CTB/McGraw-Hill had to say about the standard errors of the cut scores calculated by the standard setting committees:

> **The Standard Error of the Mean of the Standard Setting Participants' Cut Scores**
>
> Each standard setting committee's recommended cut score is computed as the mean of the participants' individual recommended cut scores. It is assumed that the participants are a sample selected from the pool of all possible participants, and if a different sample of participants had been selected, it is likely that a somewhat different cut score would result. The standard error of the cutscore (SE) is an estimate of the stability of the cut score across different standard setting committees (of the same size) randomly selected from the pool of available participants.
>
> The variability of the participants' individual cut score recommendations can be used to estimate the SE. **In general, we can assume that the cut score that would be set if the entire population of qualified participants took part in the standard setting has 68 percent likelihood of being in the interval within plus or minus 1 SE of the recommended cut score and has a 95 percent likelihood of being in the interval within plus of minus 1.96 SEs of the recommended cut score.** (*Emphasis added*.) (Lewis, 2001, p. C1-2)

These standard errors establish "confidence intervals" or bands around the cut scores. Another way to examine the consistency between the bookmark and borderline methods is to examine how close the borderline cut scores fall to

Figure 3

A comparison of the Borderline and Bookmark Methods for Determining Cut Scores

	Failure line (Basic/Below Basic)		Basic/Proficient line	
	Comparison	**Interpretation**	**Comparison**	**Interpretation**
5th Math	Borderline ***higher*** by more than 2 standard errors	Cut score higher under borderline method	Borderline ***higher*** by more than 1 standard error	Borderline Score within range likely set by bookmark
8th Math	Borderline (***higher)*** by more than 1 standard error	Borderline Score within range likely set by bookmark	Borderline ***lower*** by more than 2 standard errors	Cut score lower under borderline method
11th Math	Borderline ***lower*** by more than 2 standard errors	Cut score lower under borderline method	Borderline ***lower*** by a standard error	Borderline Score within range likely set by bookmark
5th Reading	Borderline ***higher*** by more than 3 standard errors	Cut score higher under borderline method	Borderline ***lower*** by more than 1 standard error	Borderline Score within range likely set by bookmark
8th Reading	Borderline ***higher*** by more than 3 standard errors	Cut score higher under borderline method	Borderline ***higher*** by a standard error	Borderline Score within range likely set by bookmark
11th Reading	Borderline ***higher*** by more than 3 standard errors	Cut score higher under borderline method	Borderline ***higher*** by more than 3 standard errors	Cut score higher under borderline method

Conclusions:

1. Of the six cut scores establishing the **pass-fail line**, only the Math 8 borderline score had a more than 5 percent likelihood of falling within the range likely to be set by any panel using the bookmark method.

2. It appears that the two standard setting methods produced fairly consistent results regarding the lines establishing **proficiency**.

the confidence intervals surrounding the cut scores set by the bookmark committees. Not surprisingly, we see a similar pattern of results.

When looking at the line between **basic and proficient**, we see that the borderline method produced cut scores:

- More than one standard error **above** the Math 5 bookmark cut (i.e., making it **harder** to achieve a proficient score under the borderline method)**.**
- More than two standard errors **below** the Math 8 bookmark cut (i.e., making it **easier** to achieve a proficient score under the borderline method).
- One standard error **below** the Math 11 bookmark cut score (i.e., making it **easier** to achieve a proficient score under the borderline method).

Thus, only the Math 8 borderline score had a less than 5 percent likelihood of falling within the range likely to be set by any panel using the bookmark method. In this case, the borderline score served to lower the cut. On two of the three math tests (Math 5 and 11), the bookmark and borderline basic/proficient cut scores fall within an "acceptable" range (i.e., one indicating rough agreement in the cut scores established by the two different standard-setting methods).

In the case of the line between basic and proficient on the Reading tests, the borderline score:

- Fell over one standard error **below** on the Reading 8 bookmark cut score (i.e., making it **easier** to achieve a proficient score under the borderline method).
- Fell one standard error **above** the Reading 11 bookmark cut scores (i.e., making it **harder** to achieve a proficient score under the borderline method).
- Fell over three standard errors **above** the bookmark cut score (i.e., making it **harder** to achieve a proficient score under the borderline method).

On two of the three Reading tests (Reading 8 and 11) the agreement between basic/proficient cut scores set by the different standard-setting methods is within an "acceptable" range. Thus, it appears that the two standard setting methods produced fairly consistent results regarding the lines establishing proficiency.

In contrast, comparison of the **failure line** (between basic and below basic) cut scores resulting from the two methods shows that:

- Only the Math 8 borderline cut score fell within two standard errors of the bookmark cut.

- The Math 5 borderline cut score fell more than two standard errors **above** the bookmark score.
- The Math 11 borderline cut score fell more than two standard errors **below** the bookmark score.

The most dramatic difference was in the failure cut scores for the Reading 5, 8, and 11 tests. In each case, the borderline cut score was more than three standard errors above the cut score set by the bookmark method. In other words, for Reading the borderline method produces more failures.

So of the six cut scores establishing the **pass-fail line**, only the Math 8 borderline score had a more than five percent likelihood of falling within the range likely to be set by any panel using the bookmark method.

The borderline method resulted in a higher percentage of failures on four of the six tests (one Math and three Reading). None of the Reading borderline scores had as much as a one percent likelihood of falling within the range of scores likely to be set by any panel using the bookmark method.

Research Evidence that the Borderline Method Raises Failure Rates

The tendency of the borderline standard setting to raise failure rates has been described elsewhere by the same Dr. Lewis and his co-authors. Mitzel, Lewis, Patz, and Green reported the results of a borderline groups study they conducted at a standard-setting conference (Mitzel, Lewis, Patz, and Green, 2001, pp. 273-275).

They surveyed classroom teachers before and after the bookmark standard setting and had them evaluate performance of their students' achievement and compared the evaluations with the students' actual test performance. They found little difference in the accuracy of the teachers' judgments made before and after the standard setting. However, the authors' following conclusion suggests another reason to suspect the results obtained by the borderline groups study conducted in Pennsylvania.

> What is most striking is the regularity in the functions. Across all eight grade/content combinations **students who are lower achieving tend to be underestimated**, whereas higher achieving students tend to be overestimated. Accurate estimates are, on average, obtained near the tests' mean (Ibid.).

Standard-Setting Studies Focus on Proficiency. Did Failure Line Receive Appropriate Attention?

While the CTB/McGraw-Hill Technical Report does mention the agreement of the proficient cut scores obtained by both the bookmark and borderline

studies, it does not discuss the wide disparities between the methods when setting cut scores at the lower end of the achievement spectrum.

The descriptions in both the Reading and Mathematics Technical Reports indicate the standard setting studies focused on the difference in cut scores for the proficient-basic demarcation (CTB/McGraw-Hill, 2001a, p. I-2; 2001b, p. J-2)[7].

As noted earlier, the authors of the Technical Reports did not calculate the difference in cut scores for all performance and grade levels. (Across all the Math and Reading tests the borderline procedure averaged an 8.6 percent higher failure rate per test, or a 42 percent increase in the number of students failing.)

This apparent lack of attention to the failure line is particularly problematic given the intended use of this cut score to determine those school districts designated as "failing" for Act 16 purposes and even more so now that the Schweiker Administration has proposed to expand Act 16 to include individual school buildings, adding 45 schools to the reach of the act.

Setting the Final Cuts—Important Decisions; Sparse Documentation

> "…a decision to raise or lower the recommended cut score might be based on whether the greater concern was for passing students who should have failed, or failing students who should have passed."--Dr. Daniel Lewis (Lewis 2001, pp. C1-1 to C1-2)

The Technical Reports contain no discussion or description of the methods used for establishing the final cut scores. In fact, no description was provided in any of the documentation provided by PDE in response to PSEA's Right to Know request.

As a result, we surmise the following mostly from the headings of spreadsheets PDE provided. The Department apparently decided to average

---

[7] Hambleton's seventh evaluation criterion points to another shortcoming created by the relative inattention to the failure cut score.

> (7) Were panelists explained the purposes of the educational assessment and the uses of the test scores at the beginning of the standard-setting meeting? …(A briefing on the uses of the assessment scores and the assessment itself and scoring is fundamental for panelists to set appropriate performance standards. Very different standards may result depending on the purpose of the assessment. For example, were the purpose of the assessment principally diagnostic, panelists might be expected to set fairly high standards to maximize the number of examinees who might receive assistance. A very different set of performance standards would result if the same test were being used to award high school diplomas.) (2001, p.110)

The technical documentation provided by PDE and CTB-McGraw Hill gives no indication that the bookmark and borderline groups panelists were told of the Act 16 (Education Empowerment Act) uses to which the performance levels would be put.

the cut scores resulting from the bookmark and borderline standard setting studies.

The only written rationale provided for conducting two types of standard setting studies was the single sentence contained in the documentation PDE provided to the May State Board meeting: "To maximize standard setting accuracy, the **borderline groups method** was also implemented." No precedent, source, or other rationale was found for averaging the cut scores derived from the two methods. After averaging the bookmark and borderline cut scores, PDE raised all but three cut scores by one-quarter of a standard error. (No standard error calculations were provided.)

Changes in Cut Scores Affect Thousands of Students

Based on 2000 PSSA data, the net effect of the final cut scores recommended by PDE was to move the average failure rate across the six tests to 26.2 percent as compared with a 29.2 percent average failure rate under the borderline method and a 20.6 percent failure rate under the bookmark standard setting method. The final cut scores approved by the Secretary reduced the average proportion of students reaching proficient or above to 51.7 percent from 55 percent under the bookmark method, and 53.7 percent under borderline calculations.

These seemingly small percentage differences can easily mask the impact of the Secretary's decisions. Over all six tests, when compared to the bookmark study cut scores, **the cut scores approved by the Secretary of Education decrease the number of students achieving proficient or better by over 25,000**. The PDE **cut scores increase the number of students failing by more than 42,000**. (See Appendix D.)

State Board Adopts Secretary's Recommendation; Rationale Sparse

These cut scores approved by the Secretary of Education on April 18, 2001, were recommended to and adopted by the State Board of Education at its May 2001 meeting. The only rationale offered by the Department for its adjustments has been their desire "to set the bar high," or some variation on that explanation. In the absence of a clearer explanation of the underlying rationale, the March and April 2001 adjustments made by PDE appear ad hoc.

This minimal rationale is not in accord with the ***Standards for Educational and Psychological Testing***:

> Standard 4.21
>
> When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be fully documented (AERA, APA, NCME, 1999).

Hambleton expands on this testing standard in his paper about evaluating the performance standard setting process.

> (16) Was the approach for arriving at the final performance standards clearly described and appropriate?
>
> (The approach for arriving at performance standards from the data provided by panelists may involve some complex operations. … Fitting statistical models, transforming panelist and examinee data to new scales, combining standards over sections of an assessment, and making adjustments for standard errors and or measurement errors are all common steps in arriving at performance standards.
>
> Regardless of their complexity, they need to be clearly explained, and understandable to panelists who ultimately decide on the acceptability of the performance standards. Ultimately, too, the approach used in arriving at the performance standards must be explained to boards and agencies.) (2001, p. 112)

In this instance, there is no evidence in the record that this was done by PDE in anything but the most perfunctory fashion.

It is notable that in the CTB/McGraw Hill Technical Reports provided to the Department, Dr. Lewis makes a similar statement regarding the adjustment of cut scores.

> **In sum, the literature suggests that under the proper conditions, adjustments to participants recommended cut scores can and have been made. It is clear that such adjustments should only be done with an explicit rationale, and that all information contributing to such a decision be well documented.**
>
> There are a number of reasons for adjusting the recommended cut score. First, there may be fiscal consequences of the number of students falling below a cut point. There is a conflict if students who fall below the cut point are required to receive remediation and there is a limited supply of funding for remediation. In the case, either additional funds need to be obtained or fewer students must fall below the cut score, that is, the operational cut score needs to be lowered. There is no logical reason why this cannot be publicly acknowledged, perhaps accompanied by a call for more funds.
>
> Second, there may be political consequences of the percent of students falling above or below a cut score. That is, if the recommended cut score passes (or fails) more students than is currently acceptable and/or credible, an adjustment in the cut

> score may be appropriate. ... (Emphasis added.) (Lewis, 2001, pp. C1-3 to C1-4)

Hambleton suggests additional important reasons why the standard setting agencies and boards require technical documentation.

> Technical documentation is valuable in defending the performance standards that have been set. … Often the group setting the performance standards is advisory to the agency or board that ultimately must set the standards. Technical documentation of the process is valuable information for the agency or board who must ultimately set the performance standards (Hambleton, 2001, p.108).

Again, Hambleton advises:

> *Compile validity evidence and technical documentation.* It is important not only to be systematic and thoughtful in designing and carrying out a performance standard-setting study but it is also necessary to compile validity evidence and document the work that was done and by whom (2001, p. 104).

As noted earlier, the CTB/McGraw-Hill Technical Reports were not provided to PDE until fall 2001. Thus, they could not have been available to either the Department or State Board until several months ***after*** the performance levels were established and adopted into the state regulations.

Concerns that Extent of Documentation of Two Standard-Setting Methods is Imbalanced.

The State Board's lack of information and documentation regarding the cut score setting process is one source of concern. The imbalance in the documentation eventually provided is another. The Mathematics Technical Report devotes 417 pages to documenting the bookmark standard setting studies. By contrast, it devotes 12 pages to describing the borderline groups standard setting study.

The Reading Technical Report covers the bookmark standard setting studies in 193 pages as compared to 12 pages coverage of the borderline groups study. That report refers to the "Borderline Groups Standard Setting Pilot Study" (CTB/McGraw-Hill, 2001a, p. I-1). However, it appears that the "pilot" study results were used to calculate the final Reading cut scores.

This imbalance is not trivial, given that the results of both studies were weighted equally when employed by the PDE to set actual cut scores and the final cut scores actually were substantially closer to the generally higher cut scores set by the borderline studies.[8]

[8] Although the materials PDE staff provided to the State Board May meeting mention a "Survey of Performance Expectations" designed "as a final confirmatory step before consideration by the State

Implications for State Seals: Measurement Errors and Multiple Hurdles

Different sources of error can result in setting inappropriately high or low cut scores. However, as we will see, the current method for awarding Pennsylvania state graduation seals on high school diplomas increases the likelihood that error will result in adverse consequences for the Commonwealth's graduates.

According to Michael Zieky, a standard setting pioneer at the Educational Testing Service, "The parameter being estimated in a cutscore study, however, is not the true value of the cutscore. It is, rather, the cutscore that would have been obtained if it were possible to run the particular study without any sampling error." (2001, p. 45)

The Technical Reports give no indication that the low response rates for the borderline groups Reading studies greatly increase the likelihood that the sample used was biased and that the sampling error was large.

Zieky continues:

> A different, yet equally defensible method, or a sample of participants drawn from a different, yet equally defensible population would result in a different value for the cutscore." There is general agreement now that cutscores are constructed not found. That is there is no "true" cutscore that researchers could find if only they had unlimited funding and time and could run a theoretically perfect study (Ibid.).

Lewis elaborated the tradeoff all standard-setting processes must make between different types of measurement error.

> When a student passes a test because of measurement error and not because of true ability it is called a false-positive error.
>
> In the next scenario, the opposite error occurs. In this situation, the student's obtained score is slightly below the cut score. In this case, there is some likelihood that the student's true ability is above the cut score. When a student fails a test because of measurement error and not because of true ability it is called a false-negative error.
>
> One reason that a recommended cut score would be adjusted following a standard setting would be to decrease the likelihood of a false-positive or false-negative error. Raising the recommended cut point would decrease the likelihood of false-positive errors and lowering the recommended cut point would decrease the likelihood of a false-negative error. Thus, a decision to raise or lower the recommended cut score might be

---

Board," they produced no report or compilation of the results in response to PSEA's later Right to Know request (Pennsylvania Department of Education, 2001c).

> based on whether the greater concern was for passing students who should have failed, or failing students who should have passed (2001, pp. C1-1 to C1-2).

Current plans require that in order for students to earn state seals on their high school diplomas, they will have to attain proficiency or higher on three PSSA tests: 11th grade Math, Reading, and Writing. This multiple hurdle requirement compounds identifiable problems with measurement error.

The late Lee J. Cronbach and colleagues explained this problem several years ago.

> When scores are translated into classifications, some examinees get placed in higher categories than their true performance level warrants, and some in lower categories. When a standard is applied, some examinees who deserve to fail will pass, and vice versa. The cutting score can be adjusted to make one kind of error or the other less likely. Inevitably, examinees close to the cut score will be at high risk for misclassification.
>
> *A warning on compound decision rules.* The type of analysis just made has implications for certain nonadditive passing rules. Sometimes the examinee is required to pass multiple hurdles to meet the standard (Cronbach, et al., 1995, p.10).

Multiple Hurdles Increase the Chance of False Negatives

While each hurdle may have an excellent rationale, "those who endorse [the use of multiple hurdles] **should be aware that measurement error is likely to make decisions highly fallible** (Ibid.)." (*Emphasis added*.) In other words, all that is required is for a student to receive one false negative to then be denied a diploma seal.

Senator Rhoades' recent remarks to the State Board of Education raise the question whether a single opportunity to retake the exam is an adequate corrective (Rhoades, 2002).

PSEA Estimate: PDE Cut Scores Deprive Seals to 2,500+ Students

Based on an analysis of data provided from the four school districts featured in the next section of this report (the PSEA-PASA study), we estimated that the proportion of students receiving seals drops by 2.3 percent merely because of the use of the PDE cut scores as opposed to the bookmark study cut score. If that proportion held across the state, 2,576 fewer students would receive diploma seals simply due to differences in standard-setting techniques.

Multiple Hurdle Requirement Increases this Number to 7,000-14,000

The multiple hurdle approach has an even greater impact reducing the numbers who receive seals due to the fact that some students excel in Reading or Math—but not both subjects.

Across the four districts studied:

- 64.2 percent of students scored at proficient or above in 11th grade Reading.
- 56.4 percent scored at that level in 11th grade Math.
- but only 43.2 percent scored at proficient or above on **both** tests.

The number receiving diploma seals (43.2 percent) would be much lower than the proportion achieving proficient on either test would indicate.

Statewide, far fewer than this 43.2 percent would receive a diploma seal. That's because in 2000-01 the proportions achieving proficient or higher statewide were lower than in the four districts of the PSEA-PASA study (58.1 percent in Reading 11 and 47.9 percent in Math 11)**.**

Extrapolating these numbers across the 11th grade Pennsylvania public school population, the multiple hurdles could mean that anywhere from 7,000 to 14,000 students will not earn diploma recognition for their high achievement on one of the tests. Moreover, the numbers earning seals actually will drop even further because students must also achieve proficiency on the PSSA Writing test[9].

II. Comparing Individual Performance on the PSSA Math and Reading Tests with that on Commercial Standardized Tests

Data

Four school districts agreed to participate in the PSEA-PASA study of the PSSA performance level cut scores. The selection process was nonrandom. It was based on a combined consideration of convenience and the variety of commercial standardized tests.

The test scores collected from the districts represent a census of the students who took the tests in the grades covered. (In only one case were we able to obtain scores from commercial standardized tests taken the same year as the PSSA Math and Reading tests. In the conduct of this study some districts reported that they were moving away from administering more than one set of standardized tests per school year.)

[9] This study did not examine the cut score setting process or collect any data relating to the standard setting for the PSSA Writing tests. The State Board adopted those performance standards in January 2002.

The results reported below indicate whether the commercial test was administered in the year prior or subsequent to the comparable PSSA test. The SAT tests reported were pooled, including both prior and post-PSSA dates of testing.

The data collected are suggestive, not exhaustive. The sample was not designed to be representative of the population of Pennsylvania students. Furthermore, the small numbers of students taking these tests also precludes any broad generalizations. The original purpose of the study was to see if a more comprehensive examination of the external validity of the cut scores is warranted. We believe it is.

District personnel matched the student records from the various tests and removed all identifying information prior to providing them to the author. Districts were reimbursed for the cost of this staff time.

District A is rural. Districts B, C, and D border mid-sized cities.

Explanation of graphs

The following pages contain "box and whiskers" graphs (Stata Corporation, 2001, p. 35).

The **purpose of these graphs:**

- Is to enable comparison between an individual student's percentile score on a commercial standardized test and the same student's performance on the PSSA test in the same general content area. Each graph indicates whether the commercial test was administered in the preceding, same or subsequent grade.

In these graphs the **boxes represent:**

- Students scoring from the 25^{th} to 75^{th} percentile (the inter-quartile range [IQR]) ***at each PSSA performance level*** (i.e., below basic, basic, proficient, or advanced). The midline of the box represents the commercial standardized percentile score of the median student (50^{th} percentile) scoring within the range of scores encompassed by that PSSA performance level.

  For example, in Graph 1, the midline in the box for the District A students scoring "below basic" is at 25, meaning the median student scored at the 25^{th} percentile of the nation's students who took the 4^{th} grade Terra Nova Math test the preceding year.

The **"whiskers" extending from the IQR boxes** illustrate:

- The dispersion of commercial standardized test scores and is no more than 1.5 inter-quartile ranges long. Dots beyond the end of the whiskers represent outside (extreme) values and are individually plotted.

The **width of each IQR box:**

- Is proportional to number of students it represents.

Explanation of tables

Each table corresponds to the preceding graph. They are meant to provide numerical descriptions of the relationships between an individual students percentile score on a commercial standardized test and the same student's performance on the PSSA test in the same general content area.

Student performance is grouped by the PSSA performance levels listed in the second column. Reading across the table one can see summary statistics describing how students in each performance category fared on the commercial test as well as on the PSSA test.

Graphs 1 through 3 depict the relationship between performance on the PSSA 5th grade Mathematics test, the Terra Nova 4th grade Math test (2 districts), and the 5th grade California Achievement Math Test (CAT-5). The Terra Nova and California Achievement tests are published by McGraw-Hill. Graph 4 depicts the relationship between performance on the PSSA 5th grade Mathematics test and the Comprehensive Testing Program III (CTP-III) 5th grade Math test. The latter is published by the Education Records Bureau in cooperation with the Educational Testing Service (ETS).

The first four graphs indicate the following:

- There is a strong relationship between performance on the 5th grade PSSA Mathematics test and the three commercial tests. One would expect a strong relationship in the results of two achievement tests in the same content area.
- Students scoring below basic on PSSAs approach or surpass the 50th percentile on national commercial tests, suggesting that the below basic cut score is set too high.
- Students scoring on average are performing roughly between the 25th and 38th percentiles as measured against the national samples of the Terra Novas. Below basic students in District C performed similarly around the 34th percentile on the California Achievement Test (CAT)-5, while

District D's below basic students averaged near the 43rd percentile of the CTP-III. [10]

- Students achieving the basic level on the PSSAs are clustering about the 44th to 64th percentiles of the Terra Nova, the 57th percentile of the CAT-5 and the 60th percentile of the CTP-III.
- Proficient students are averaging around the 72nd-73rd percentiles of the Terra Nova, the 76th of the CAT-5 and the 79th of the CTP-III. Advanced students are averaging around the 83rd-87th percentiles of the Terra Nova, 92nd of the CAT-5 and 94th of the CTP-III.

Graphs 5 through 8 depict the relationship between performance on the 5th grade PSSA Reading test and performance on the Reading tests of the same publishers detailed in Graphs 1-4. The pattern of results from these graphs and tables is very similar to the pattern of the first four. Again, we can see a strong relationship between performance on the 5th Grade PSSA test (Reading) and performance on the commercial tests. We can see that the PSSA 5th grade Reading performance standards classify many students as failing, when in fact, they are meet the mean on national achievement tests. The only difference is that students achieving at the different PSSA performance levels, on average, are scoring in slightly lower percentiles on the nationally normed tests.

Graphs 9 through 12 show the relationship between performance on the 8th grade PSSA Mathematics test and performance on the Terra Nova Math 7, Terra Nova Math 9, Terra Nova Math 7, and CTP-III 7th grade Math tests, respectively. The patterns again are very similar to the patterns depicted in Graphs 1 to 4.

- Students scoring below basic on the 8th grade PSSA Math test average from the 31st to 44th percentiles on these tests.
- **Students achieving at the basic level averaged scoring from the 49th to 62nd percentiles.**
- Proficient scorers averaged from the 69th to 74th percentile.
- Advanced achievers averaged from the 85th to 94th percentiles.

Graphs 13 through 16 continue the patterns described above. We can see the PSSA 8th grade Reading performance standards are also very rigorous. Again, there is a slight lowering of their average percentile on the commercial tests at below basic and basic (compared with the Math 8 results).

- Below basic PSSA performers are achieving around the 30th percentile nationally.
- Basic from the 41st to 56th.
- Proficient from the 59th to the 74th .
- Advanced from the 84th to 94th.

[10] Compare this performance with the sub-25th percentile performance of students in the "failure" category of the Massachusetts [MCAS] state 4th grade math test (See Gong, 1999).

Graphs containing results for the 11th grade PSSA tests differ from the previous sets in a number of important respects. Graphs 18 and 20 contained pooled SAT I (Math and Verbal Reasoning) results from Districts A, B, and C. (The SAT is developed by ETS for the College Board.) The horizontal lines across the SAT graphs correspond with scaled scores marking the 25th, 50th, and 75th percentiles of the SAT I distributions. Note that the position of these lines changes slightly from the Math to Verbal tests, reflecting differences in the shape of their score distributions (see College Entrance Examination Board, 2001, p. 9). Graph D again contains a single district's CTP-III 10th grade Math and Reading results.

As is the case with all previous tests, the 11th grade PSSA Math and Reading performance standards establish cut scores that see most "basic" performers placed among the top half of the nation on the commercial tests. In District D, the average CTP-III performance of students scoring at each PSSA performance level was very similar across the Math (Graph 17) and Reading (Graph 19) tests.

- Below basic performers averaged in the low 40s.
- Basic performers averaged at the 61st percentile.
- Proficient in the low 70s.
- Advanced achievers around the 90th percentile.

Note that the SAT comparison group is comprised of college-bound students. The SAT I Math Reasoning results again show a strong relationship between the PSSA performance levels and a commercial test. However, the below basic category (perhaps because of a substantially lower sample size) is much less distinct. Graph 18 shows that the dispersion of below basic performers is so wide as to substantially overlap and encompass the range of the basic performers. Below basic performers are scoring, on average, in the bottom quartile of their college-bound peers. Basic performers are scoring in the next quartile, while proficient performers are scoring above the 50th percentile. Advanced achievers on the PSSA Math 11 test are scoring in the top 25 percent of the SAT I math distribution.

Graph 20 provides additional evidence that the failure cut score on the 11th grade PSSA Reading test bears re-examination. Looking at the graph and table, we can see that ***below basic PSSA performers averaged slightly higher SAT performance than those classified as basic***. The high degree of overlap between the SAT performance of students achieving at these two performance levels no doubt is susceptible to multiple interpretations. We believe the prudent course would be to gather further evidence concerning the validity of the cut score and make adjustments if indicated.

Graph 1

District A: Comparison of Individual's Scores on Terra Nova Math 4 and PSSA Math 5

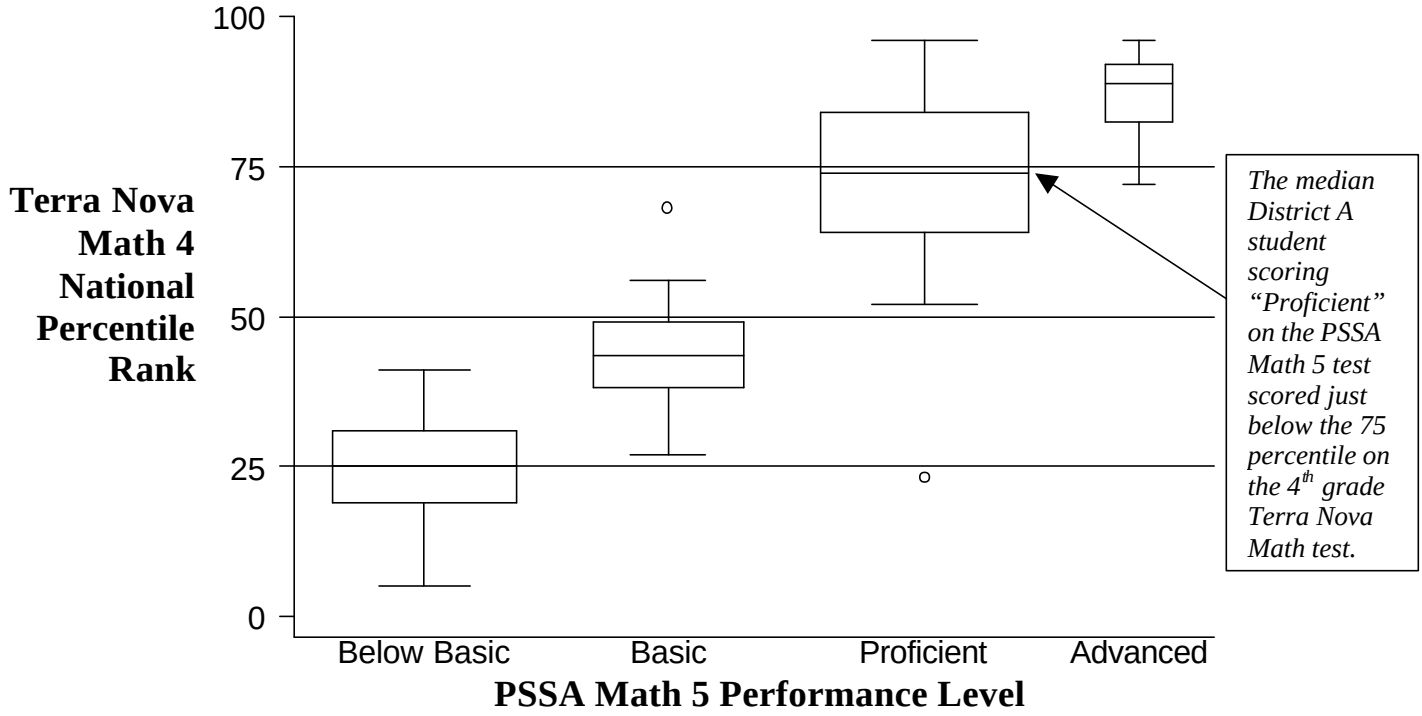
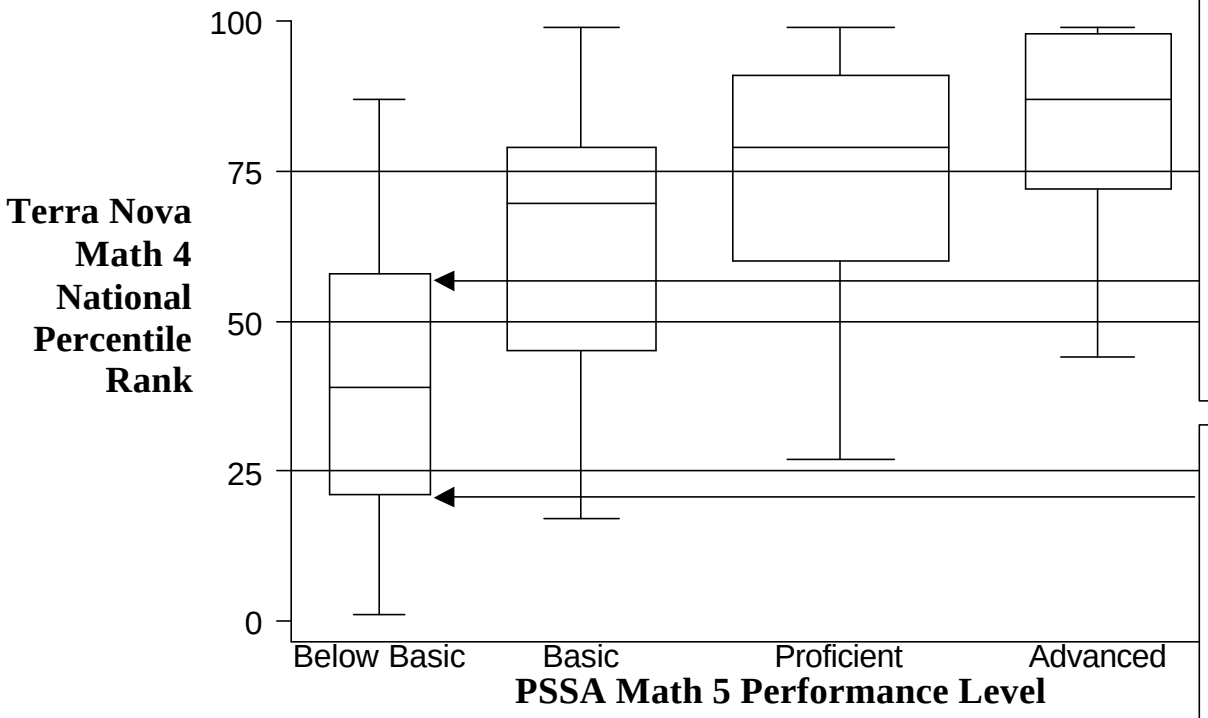


Table 1

District A	Terra Nova Math 4 Percentile			PSSA Scaled Score		
	N	Mean	Std. Dev.	N	Mean	Std. Dev.
PSSA Math 5						
Below Basic	**22**	**24.8**	**9.5**	**24**	**1070**	**66.3**
Basic	**18**	**44.2**	**9.4**	**18**	**1226**	**41.0**
Proficient	**25**	**72.4**	**16.2**	**25**	**1365**	**32.0**
Advanced	**8**	**86.9**	**7.8**	**8**	**1591**	**80.2**

Graph 2

District B: Comparison of Individual's Scores on Terra Nova Math 4 and PSSA Math 5



The District B student at the 75th percentile of the District's students scoring "Below Basic" on the PSSA Math 5 test, scored well above the 50th percentile on the Terra Nova. Math 4 test.

The District B student at the 25th percentile of the District's students scoring "Below Basic" on the PSSA Math 5 test, scored just below the 25th percentile on the Terra Nova Math 4 test.

Table 2

District B

	Terra Nova Math 4 Percentile			PSSA Scaled Score		
PSSA Math 5	**N**	**Mean**	**Std. Dev.**	**N**	**Mean**	**Std. Dev.**
Below Basic	**37**	**37.5**	**21.5**	**48**	**1059**	**75.8**
Basic	**54**	**64.1**	**21.8**	**64**	**1235**	**40.0**
Proficient	**78**	**73.2**	**20.2**	**98**	**1374**	**42.9**
Advanced	**53**	**83.1**	**16.1**	**58**	**1573**	**123.1**

Graph 3

District C: Comparison of Individual's Scores on California Achievement Test (CAT-5) Math
and PSSA Math 5

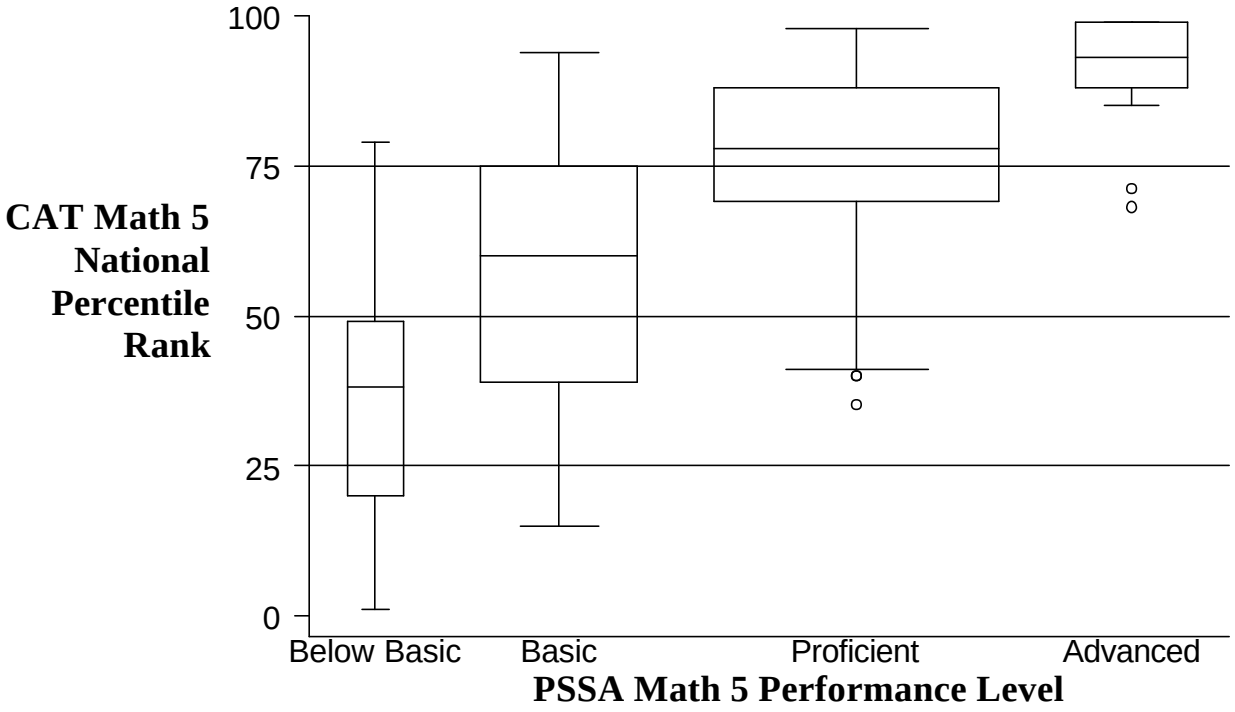


Table 3

District C	CAT Math 5 Percentile			PSSA Scaled Score		
PSSA Math 5	N	Mean	Std. Dev.	N	Mean	Std. Dev.
Below Basic	13	34.1	21.1	27	1016	114.5
Basic	35	56.7	20.5	37	1227	39.7
Proficient	64	76.0	15.6	64	1368	43.0
Advanced	25	92.0	8.3	25	1550	108.2

Graph 4

District D: Comparison of Individual's Scores on the Education Records Bureau (CTP III) 5th Grade Math and PSSA Math 5

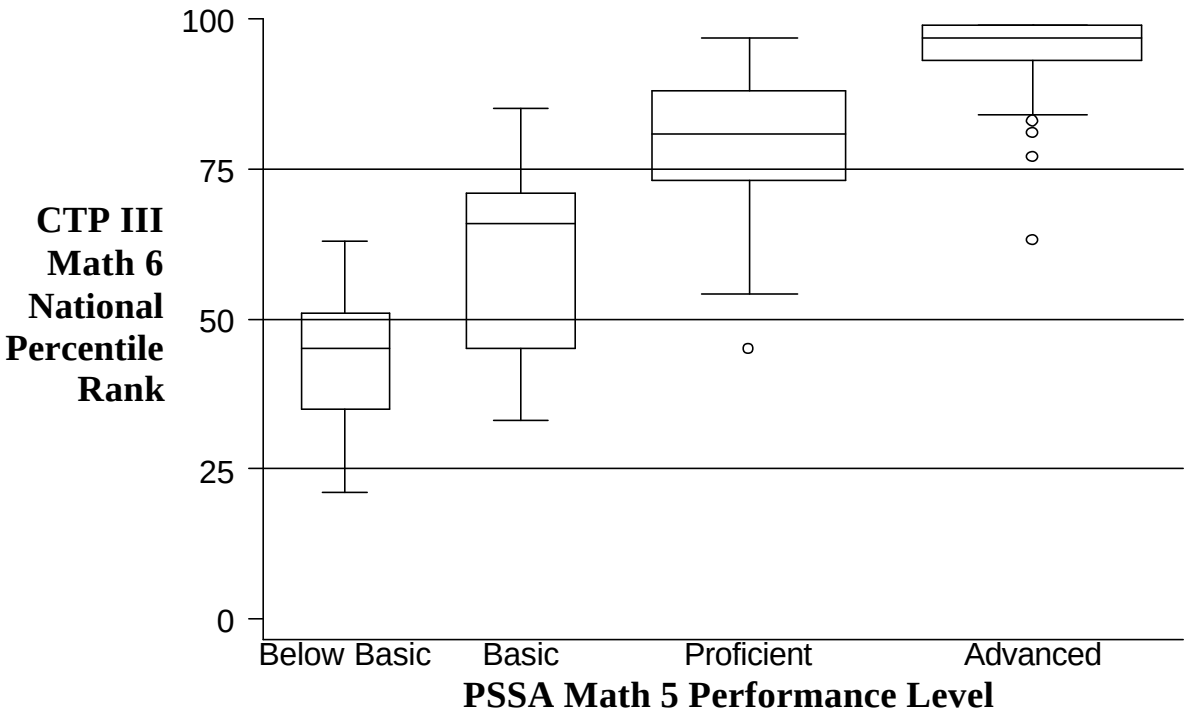


Table 4

District D		ERB-CTP III Math 6 Percentile			PSSA Scaled Score		
		N	Mean	Std. Dev.	N	Mean	Std. Dev.
PSSA Math 5							
Below Basic		17	42.8	13.3	20	1036	252.6
Basic		21	60.0	16.1	26	1246	37.9
Proficient		37	79.2	12	41	1393	48.1
Advanced		42	94.0	7.6	43	1623	132.4

Graph 5

District A: Comparison of Individual's Scores on Terra Nova Reading 4 and PSSA Reading 5

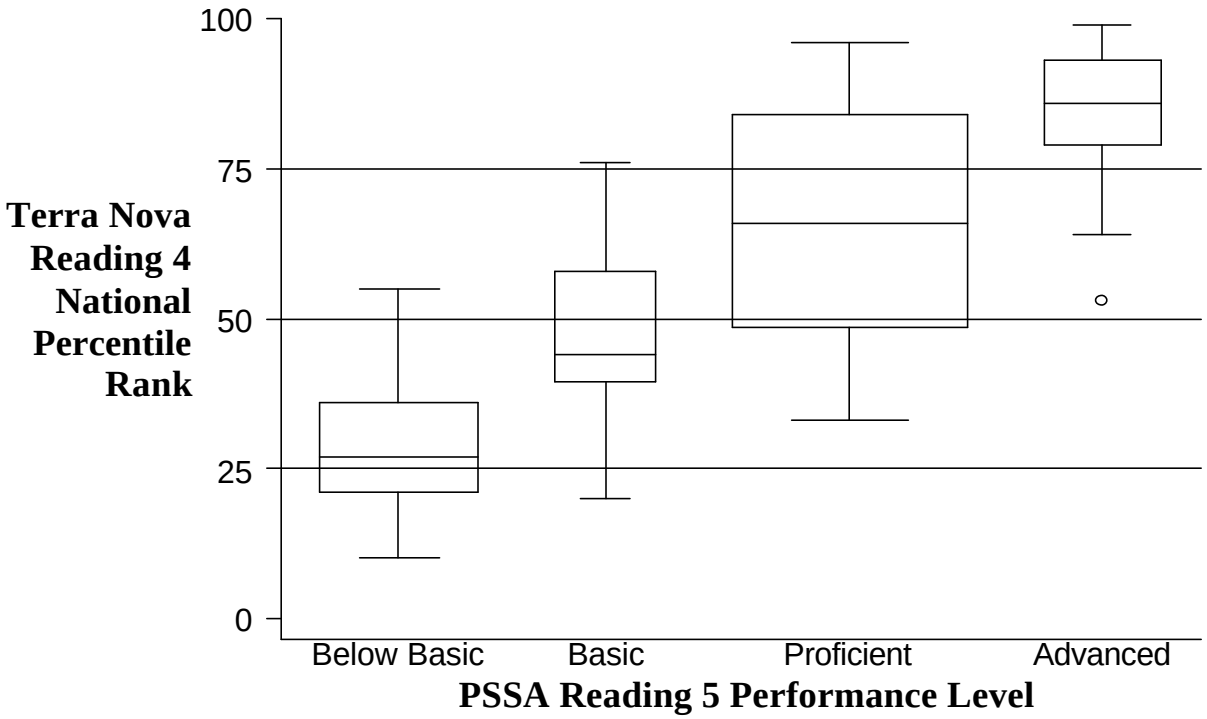


Table 5

District A		Terra Nova Reading 4 Percentile			PSSA Scaled Score		
		N	Mean	Std. Dev.	N	Mean	Std. Dev.
	PSSA Reading 5						
	Below Basic	**19**	**28.5**	**11.4**	**21**	**1028**	**102.8**
	Basic	**12**	**46.0**	**15.5**	**12**	**1230**	**28.7**
	Proficient	**28**	**65.1**	**20.2**	**28**	**1386**	**55.2**
	Advanced	**14**	**84.2**	**13.2**	**14**	**1563**	**77.4**

Graph 6

District B: Comparison of Individual's Scores on Terra Nova Reading 4 and PSSA Reading 5

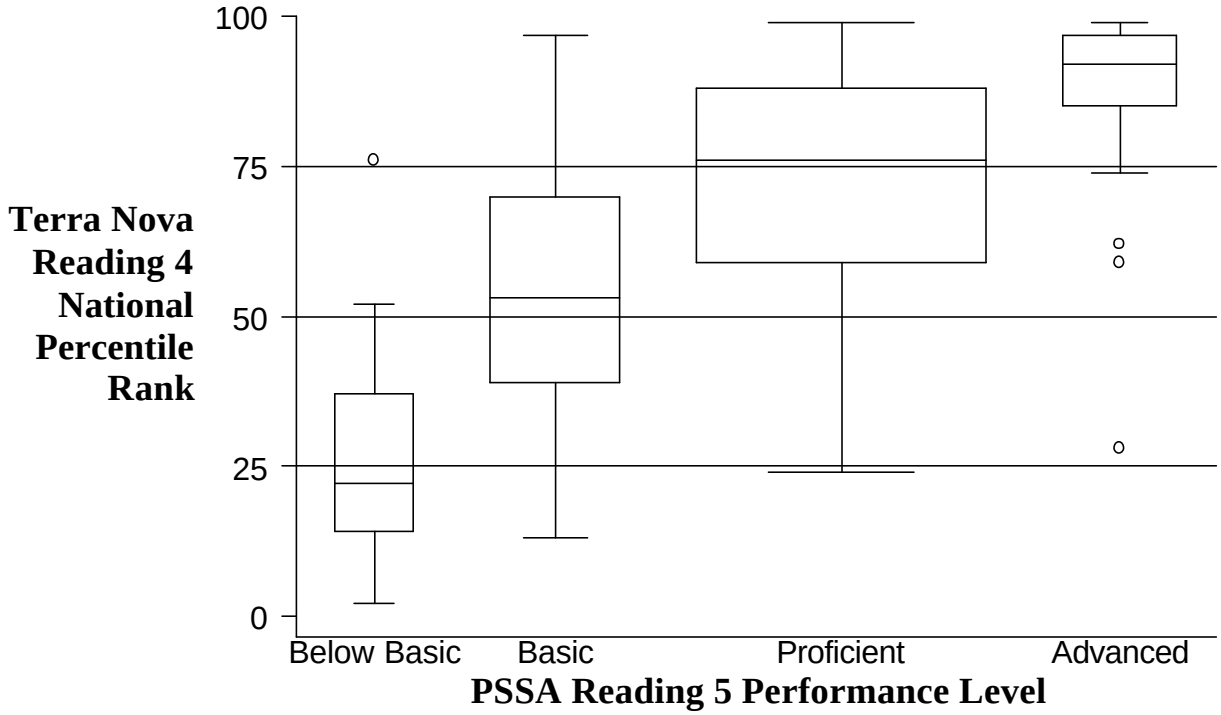


Table 6

District B	Terra Nova Reading 4 Percentile			PSSA Scaled Score		
	N	Mean	Std. Dev.	N	Mean	Std. Dev.
PSSA Reading 5						
Below Basic	**29**	**25.8**	**17.5**	**44**	**1034**	**93.0**
Basic	**47**	**53.1**	**20.7**	**56**	**1233**	**44.3**
Proficient	**105**	**72.2**	**19.1**	**119**	**1394**	**56.2**
Advanced	**41**	**88.12**	**13.7**	**49**	**1575**	**87.0**

Graph 7

District C: Comparison of Individual's Scores on
California Achievement Test (CAT-5) Reading and PSSA Reading 5

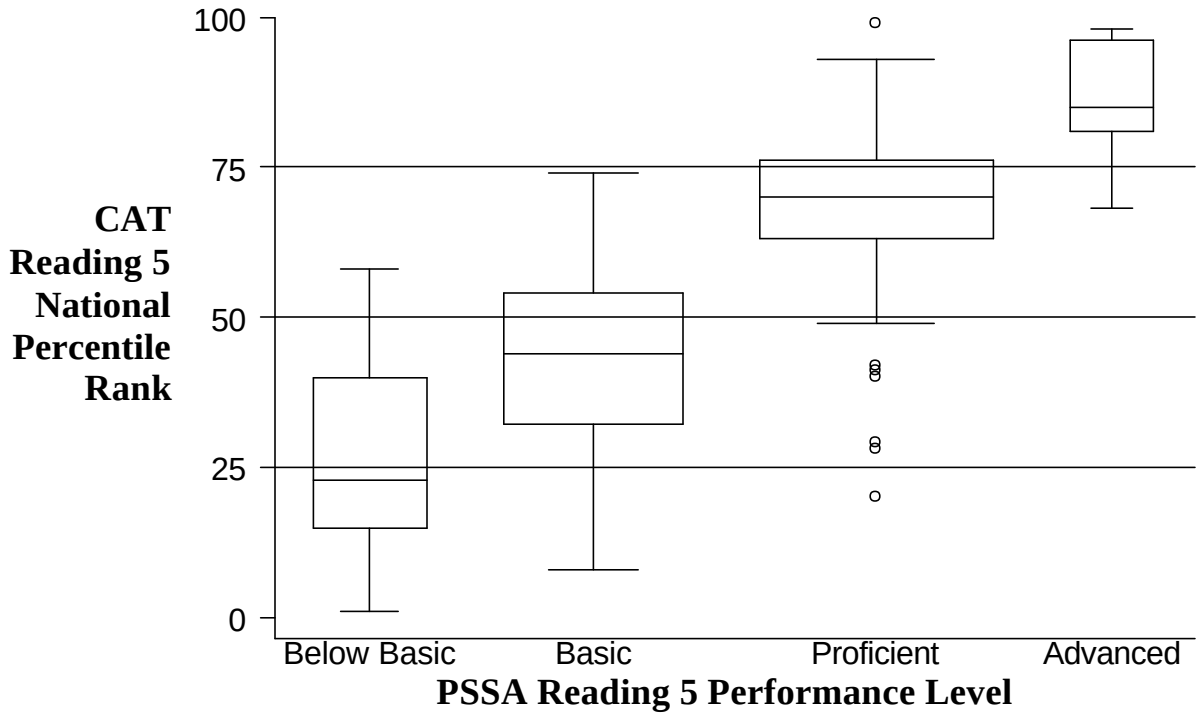


Table 7

District C	CAT Reading 5 Percentile			PSSA Scaled Score		
	N	Mean	Std. Dev.	N	Mean	Std. Dev.
PSSA Reading 5						
Below Basic	**26**	**26.5**	**15.6**	**39**	**1000**	**120.3**
Basic	**41**	**43.2**	**16.3**	**42**	**1233**	**39.8**
Proficient	**53**	**68.6**	**16.4**	**53**	**1387**	**58.0**
Advanced	**19**	**86.2**	**8.7**	**19**	**1563**	**77.2**

Graph 8

District D: Comparison of Individual's Scores on the Education Records Bureau (CTP-III) 6th Grade Reading and PSSA Reading 5

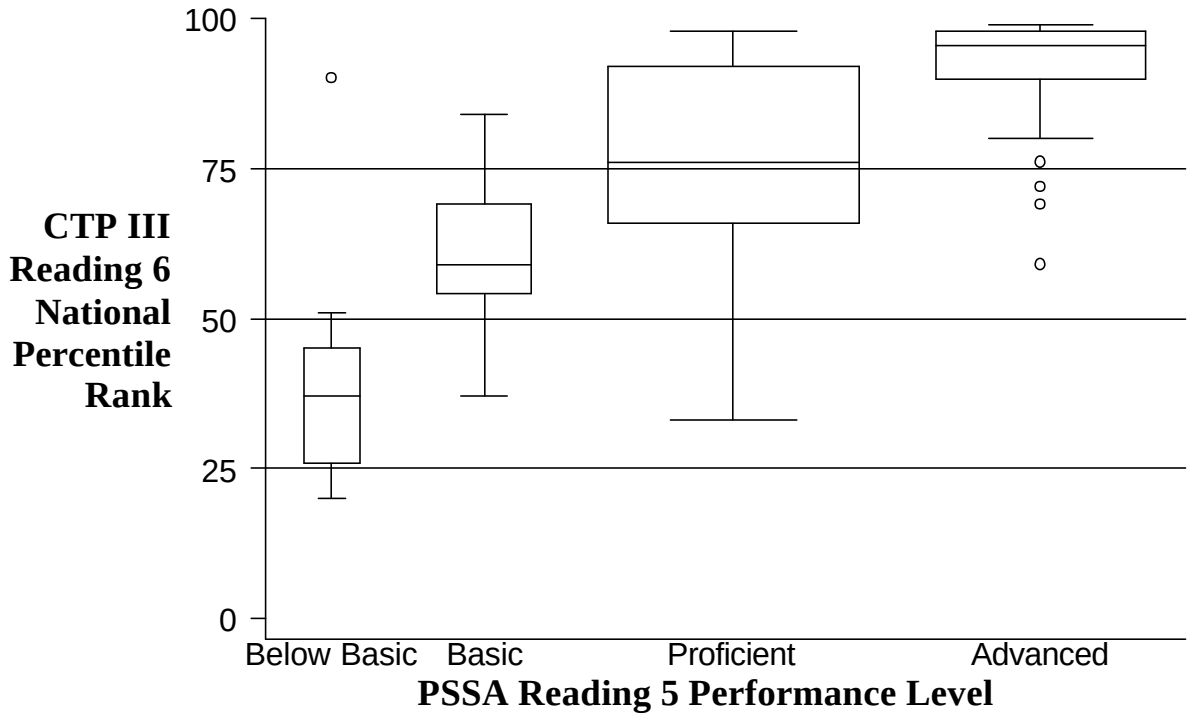


Table 8

District D		ERB-CTP III Reading 6 Percentile			PSSA Scaled Score		
		N	Mean	Std. Dev.	N	Mean	Std. Dev.
PSSA Reading 5							
Below Basic		11	40.5	19.3	14	995	297
Basic		18	60.2	12.8	22	1242	45.4
Proficient		48	77.1	15.5	51	1388	44.9
Advanced		40	91.8	9.5	43	1577	70.5

Graph 9

District A: Comparison of Individual's Scores on Terra Nova Math 7 and PSSA Math 8

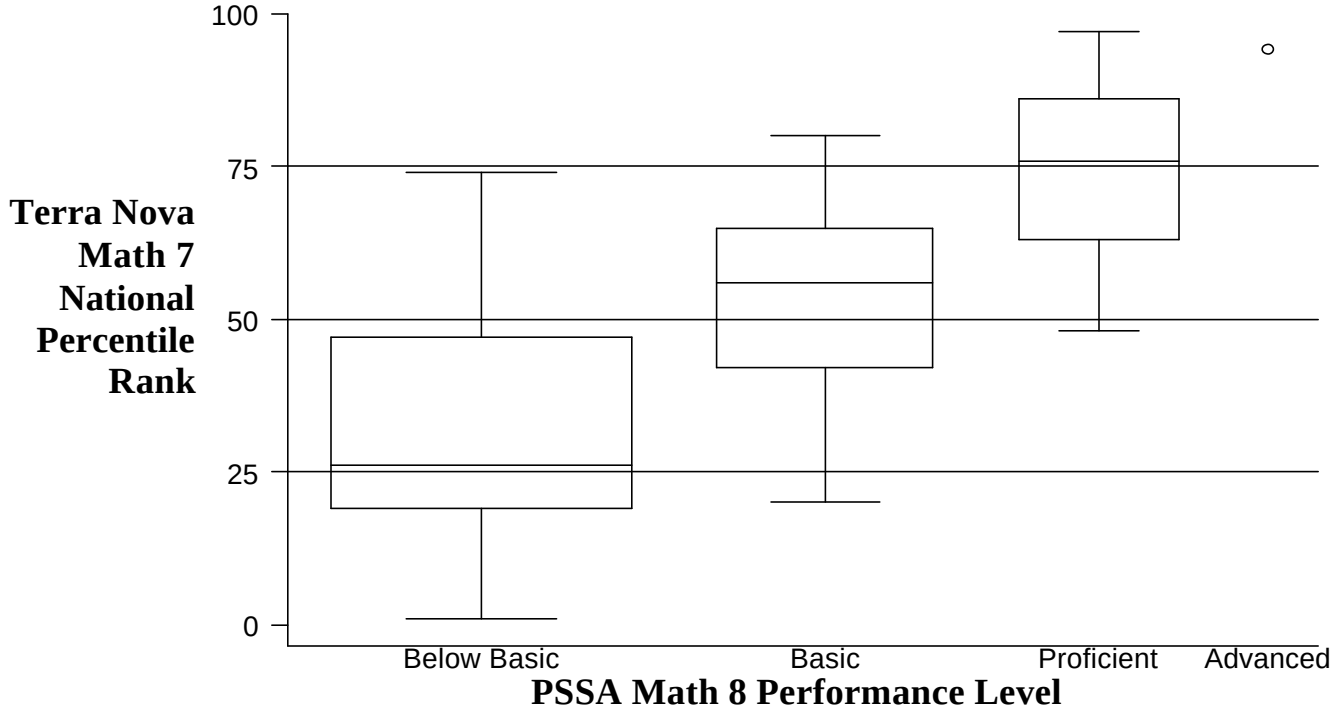


Table 9

District A	Terra Nova Math 7 Percentile			PSSA Scaled Score		
	N	Mean	Std. Dev.	N	Mean	Std. Dev.
PSSA Math 8						
Below Basic	**32**	**31.8**	**19.4**	**37**	**1080**	**62.1**
Basic	**23**	**53.2**	**17.5**	**26**	**1235**	**35.7**
Proficient	**17**	**74.4**	**15.0**	**17**	**1394**	**52.1**
Advanced	**1**	**94**	**N/A**	**1**	**1637**	**N/A**

Graph 10

District B: Comparison of Individual's Scores on Terra Nova Math 9 and PSSA Math 8

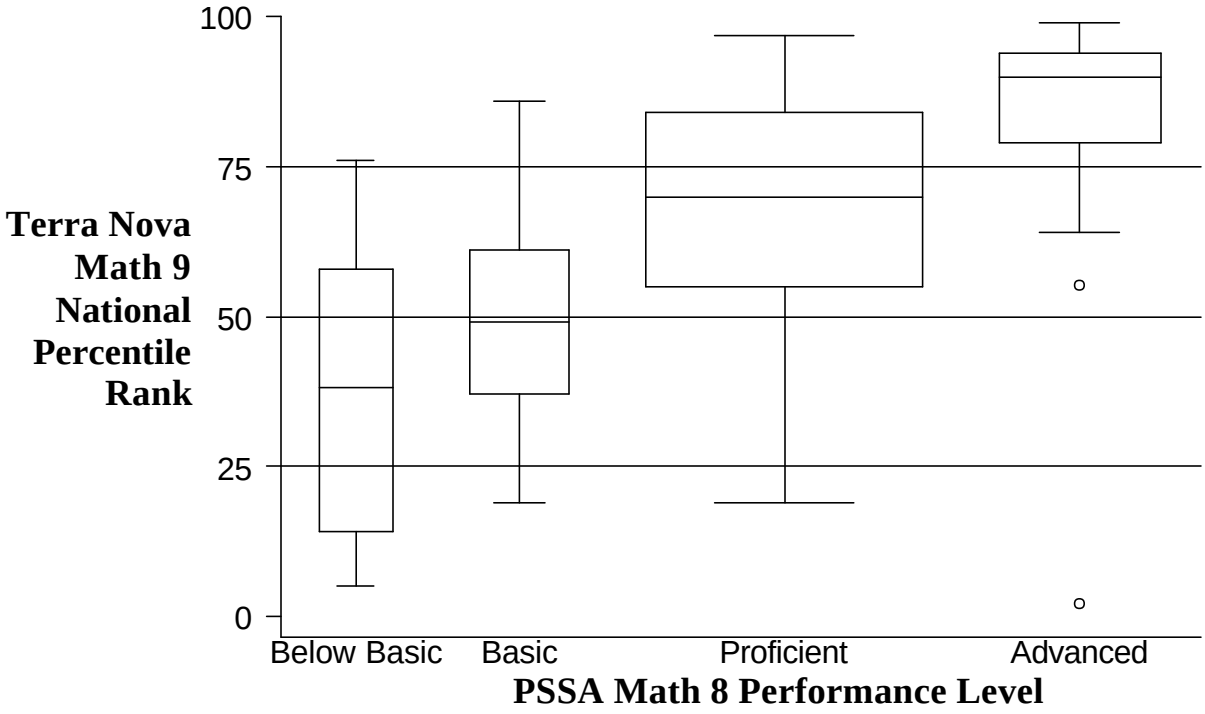


Table 10

District B		Terra Nova Math 9 Percentile			PSSA Scaled Score		
	PSSA Math 8	N	Mean	Std. Dev.	N	Mean	Std. Dev.
	Below Basic	31	36.7	22.9	41	1065	87.9
	Basic	42	49.1	16.8	47	1238	33.9
	Proficient	115	69.0	18.9	126	1412	64.3
	Advanced	67	85.2	14.4	69	1621	83.0

Graph 11

District B: Comparison of Individual's Scores on Terra Nova Math 7 and PSSA Math 8

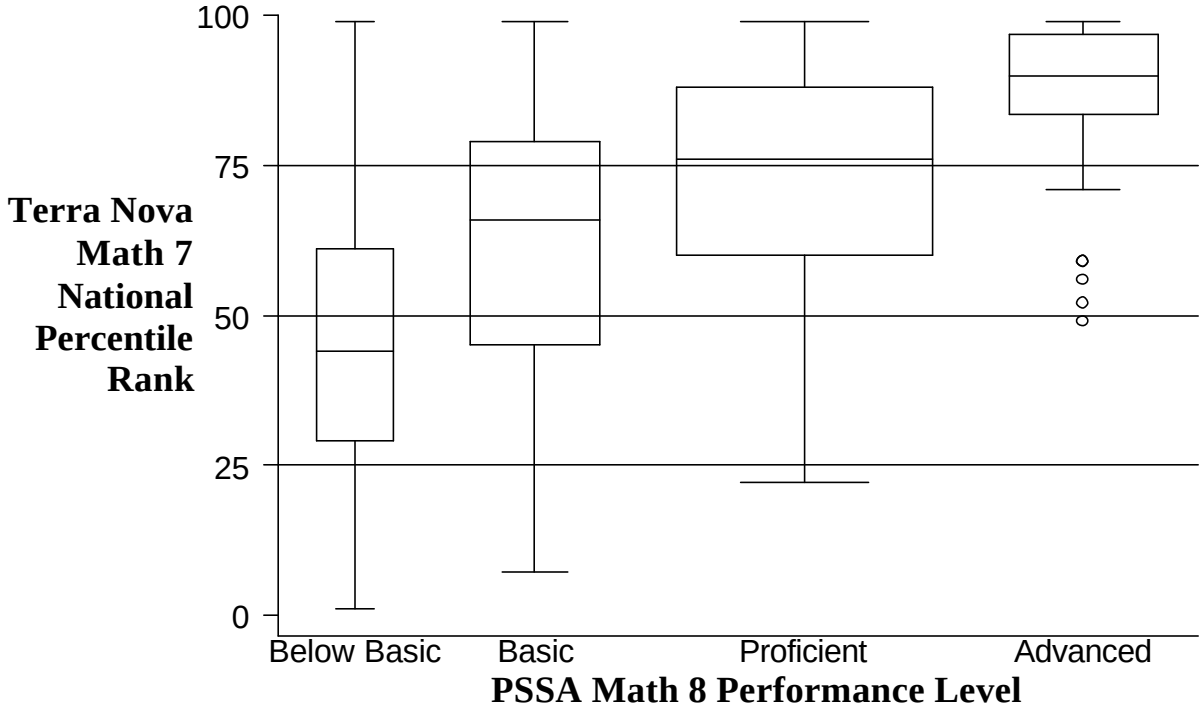


Table 11

District B		Terra Nova Math 7 Percentile			PSSA Scaled Score		
	PSSA Math 8	N	Mean	Std. Dev.	N	Mean	Std. Dev.
	Below Basic	31	44.1	21.5	39	1060	68.1
	Basic	53	62.2	22.6	60	1238	38.4
	Proficient	103	73.1	18.8	111	1408	57.8
	Advanced	60	86.8	12.5	62	1623	90.5

Graph 12

District D: Comparison of Individual's Scores on the Education Records Bureau (CTP III) 7th Grade Math and PSSA Math 8

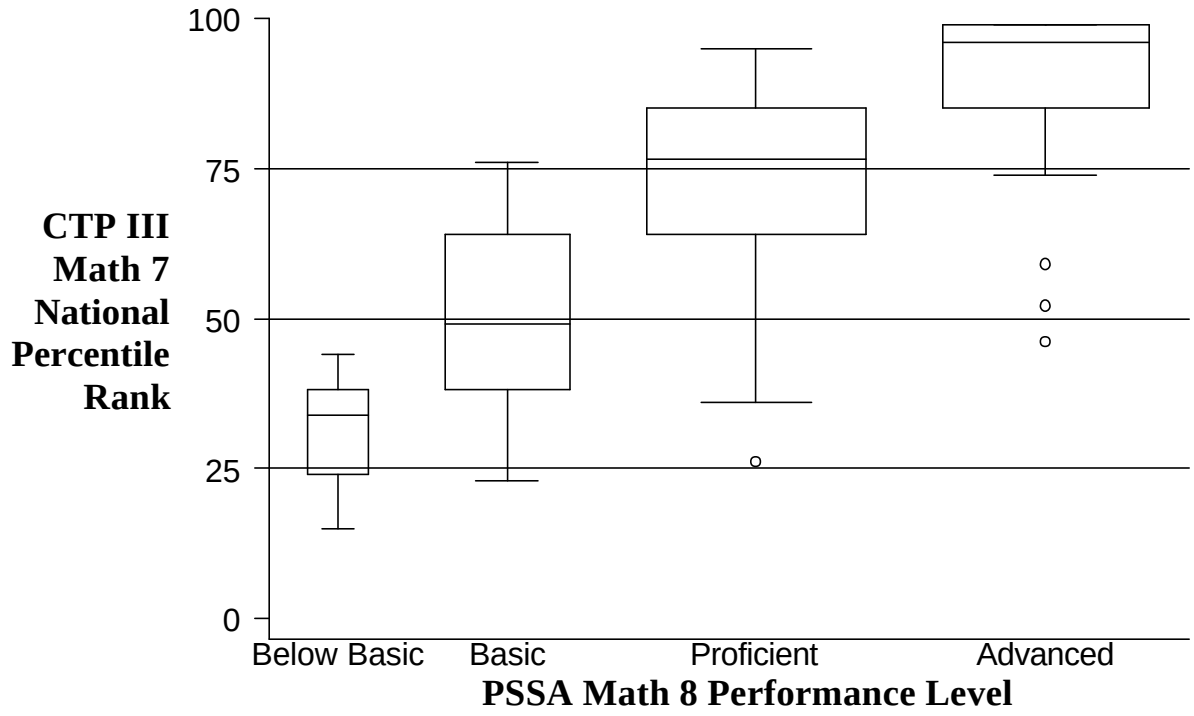


Table 12

District D		ERB-CTP III Math 7 Percentile			PSSA Scaled Score		
		N	Mean	Std. Dev.	N	Mean	Std. Dev.
	PSSA Math 8						
	Below Basic	**13**	**31.0**	**9.4**	**19**	**1004**	**251.0**
	Basic	**26**	**49.7**	**15.5**	**28**	**1247**	**31.6**
	Proficient	**46**	**72.2**	**15.8**	**47**	**1407**	**59.0**
	Advanced	**43**	**90.4**	**12.7**	**48**	**1679**	**170.7**

Graph 13

District A: Comparison of Individual's Scores on Terra Nova Reading 7 and PSSA Reading 8

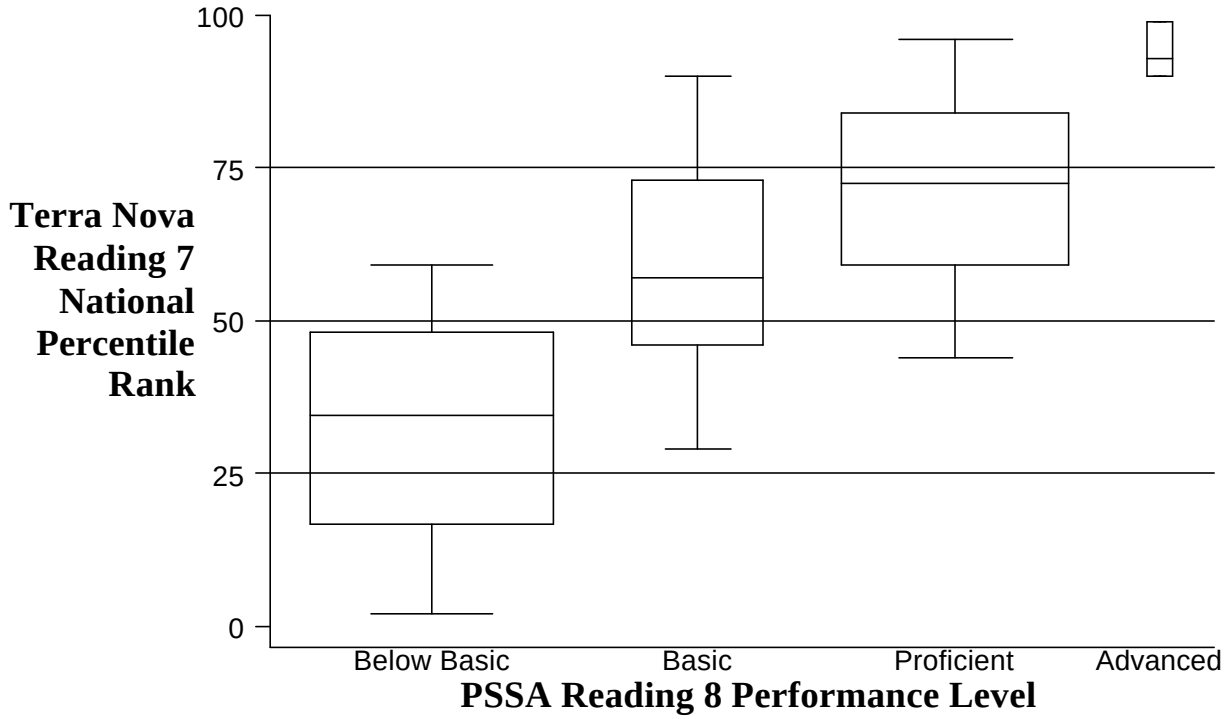


Table 13

District A	Terra Nova Reading 7 Percentile			PSSA Scaled Score		
	N	Mean	Std. Dev.	N	Mean	Std. Dev.
PSSA Reading 8						
Below Basic	**32**	**31.8**	**19.4**	**37**	**1080**	**62.1**
Basic	**23**	**53.2**	**17.5**	**26**	**1235**	**35.7**
Proficient	**17**	**74.4**	**15.0**	**17**	**1394**	**52.1**
Advanced	**1**	**94**	**N/A**	**1**	**1637**	**N/A**

Graph 14

District B: Comparison of Individual's Scores on Terra Nova Reading 9 and PSSA Reading 8

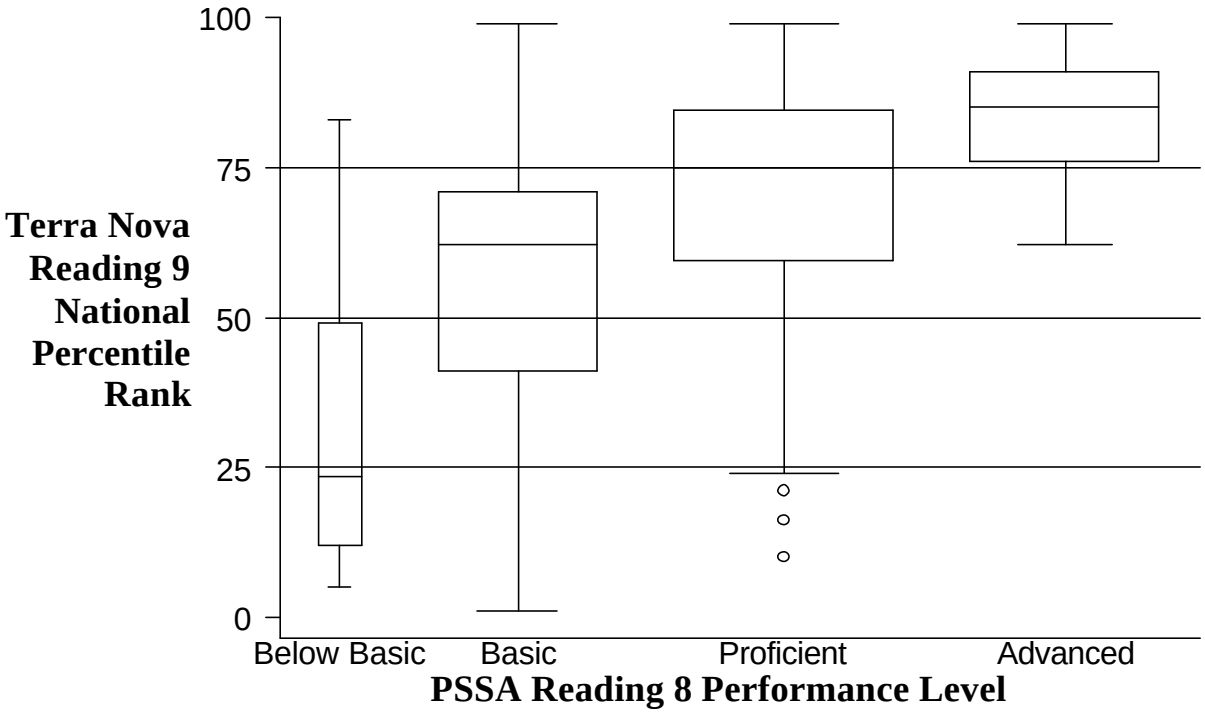


Table 14

District B		Terra Nova Reading 9 Percentile			PSSA Scaled Score		
		N	Mean	Std. Dev.	N	Mean	Std. Dev.
PSSA Reading 8							
Below Basic		18	31.3	24.8	21	1072	46.9
Basic		67	56.2	23.1	69	1211	42.2
Proficient		92	70.8	18.3	94	1380	58.4
Advanced		79	84.0	9.7	80	1638	147.4

Graph 15

District B: Comparison of Individual's Scores on Terra Nova Reading 7 and PSSA Reading 8

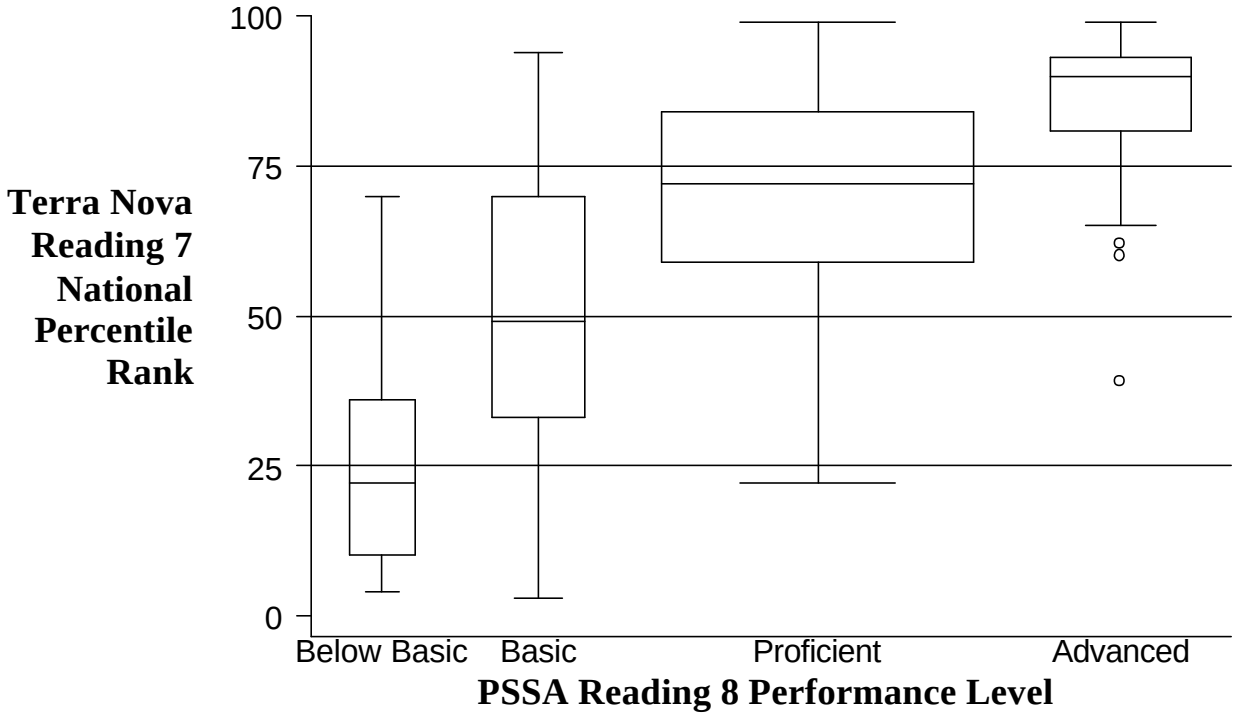


Table 15

District B	Terra Nova Reading 7 Percentile			PSSA Scaled Score		
	N	Mean	Std. Dev.	N	Mean	Std. Dev.
PSSA Reading 8						
Below Basic	**27**	**26.7**	**18.4**	**34**	**979**	**93.7**
Basic	**38**	**50.0**	**22.1**	**46**	**1216**	**44.7**
Proficient	**126**	**70.5**	**17.3**	**135**	**1395**	**58.1**
Advanced	**57**	**85.7**	**12.1**	**60**	**1575**	**63.9**

Graph 16

District D: Comparison of Individual's Scores on the Education Records Bureau (CTP-III) 7th Grade Reading and PSSA Reading 8

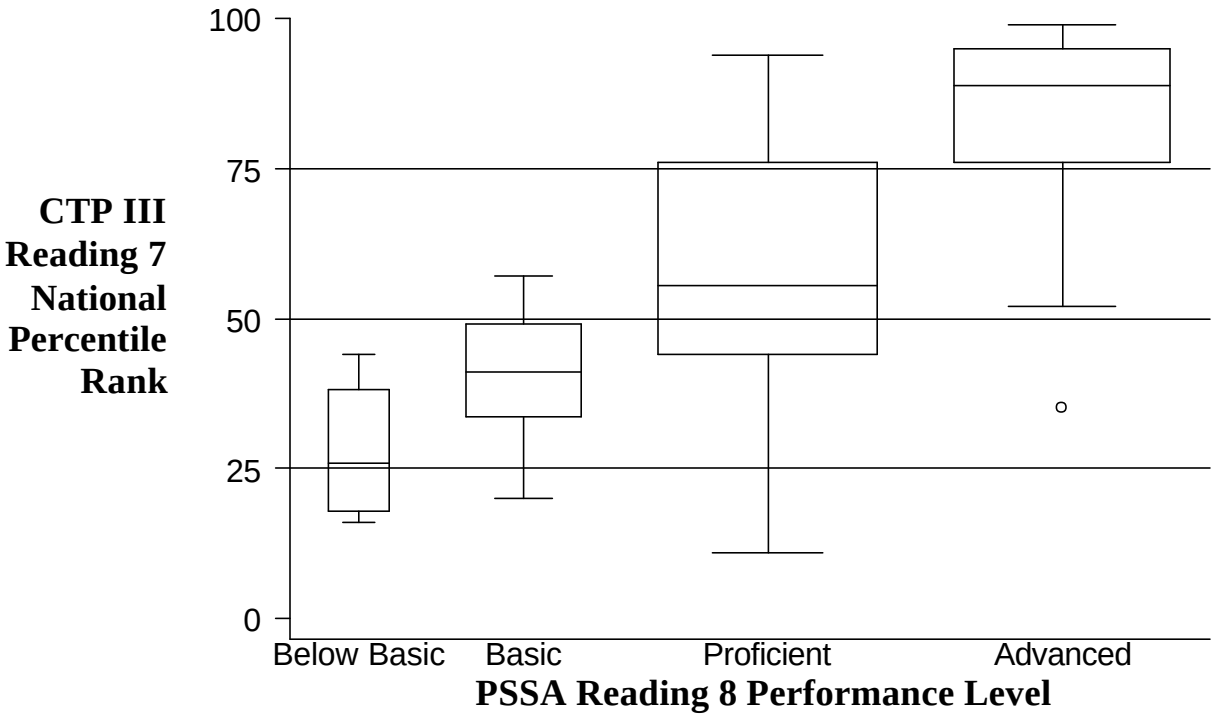


Table 16

District D		ERB-CTP III Reading 7 Percentile			PSSA Scaled Score		
		N	Mean	Std. Dev.	N	Mean	Std. Dev.
PSSA Reading 8							
Below Basic		13	28.1	10.2	17	967	260.9
Basic		24	40.8	10.3	26	1215	38.3
Proficient		46	59.0	21.0	52	1385	64.3
Advanced		45	83.5	15.7	47	1604	82.9

Graph 17

District D: Comparison of Individual's Scores on the Education Records Bureau (CTP III) 10th Grade Math and PSSA Math 11

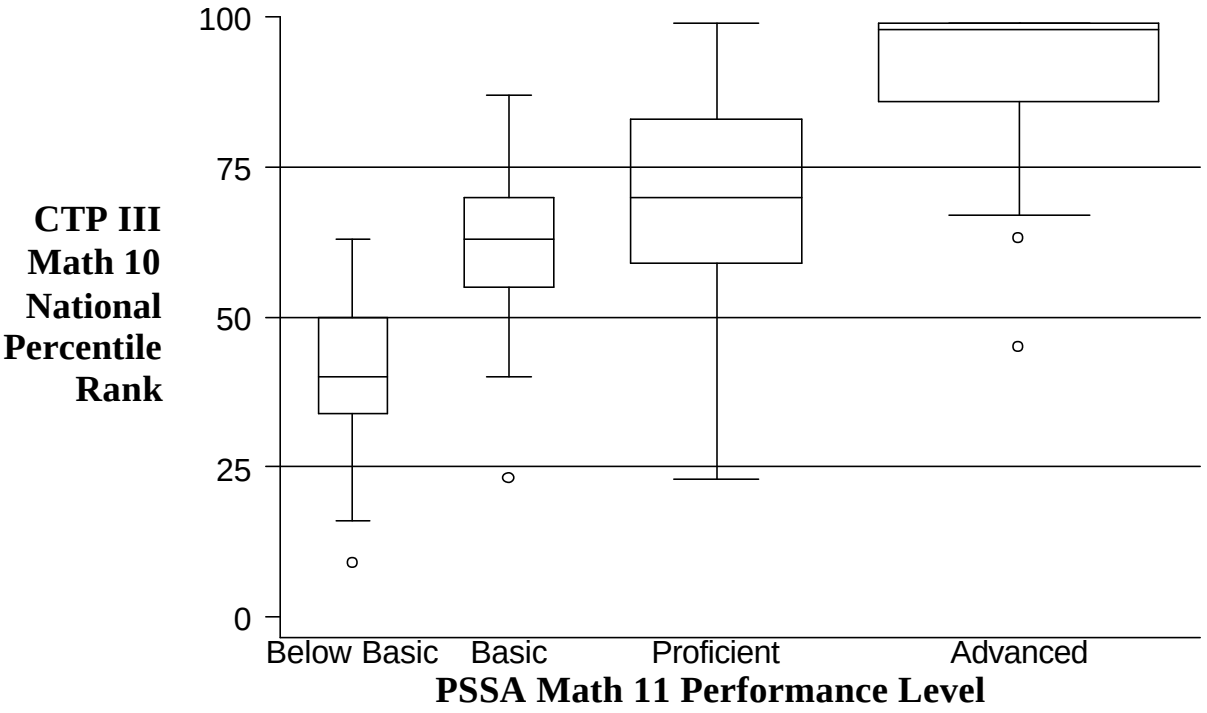


Table 17

District D	ERB-CTP III Math 10 Percentile			PSSA Scaled Score		
	N	Mean	Std. Dev.	N	Mean	Std. Dev.
PSSA Math 11						
Below Basic	**14**	**40.2**	**15.3**	**19**	**960**	**347.6**
Basic	**18**	**61.2**	**15.0**	**20**	**1244**	**38.4**
Proficient	**34**	**70.9**	**16.0**	**37**	**1397**	**39.7**
Advanced	**56**	**91.8**	**11.2**	**62**	**1654**	**142.5**

Graph 18

Districts A, B, &C: Comparison of Individual's Scores on the SAT I Reasoning Test (Math) and PSSA Math 11

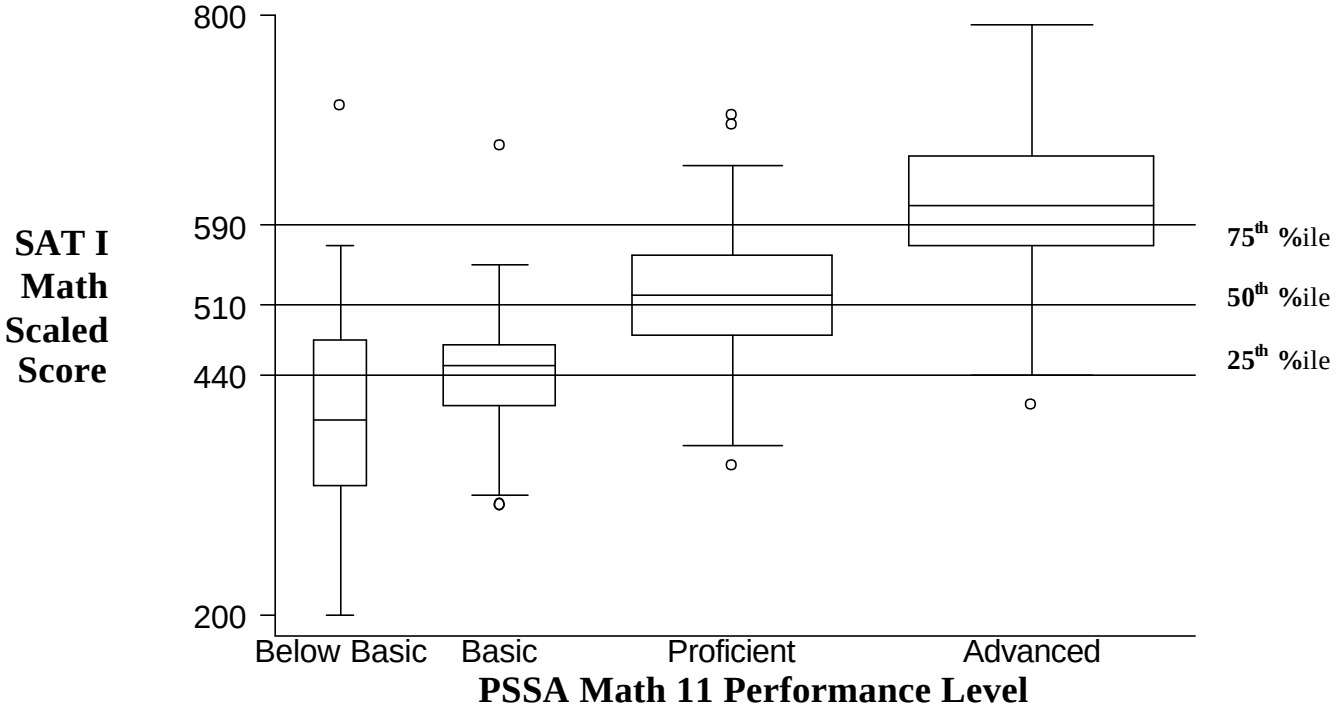


Table 18

Districts A, B, & C		SAT I Math Scaled Score	
	N	**Mean**	**Std. Dev.**
PSSA Math 11			
Below Basic	**36**	**404**	**100**
Basic	**75**	**446**	**57.5**
Proficient	**133**	**517**	**60.1**
Advanced	**163**	**617**	**69.8**

Graph 19

District D: Comparison of Individual's Scores on the Education Records Bureau (CTP III) 10th Grade Reading and PSSA Reading 11

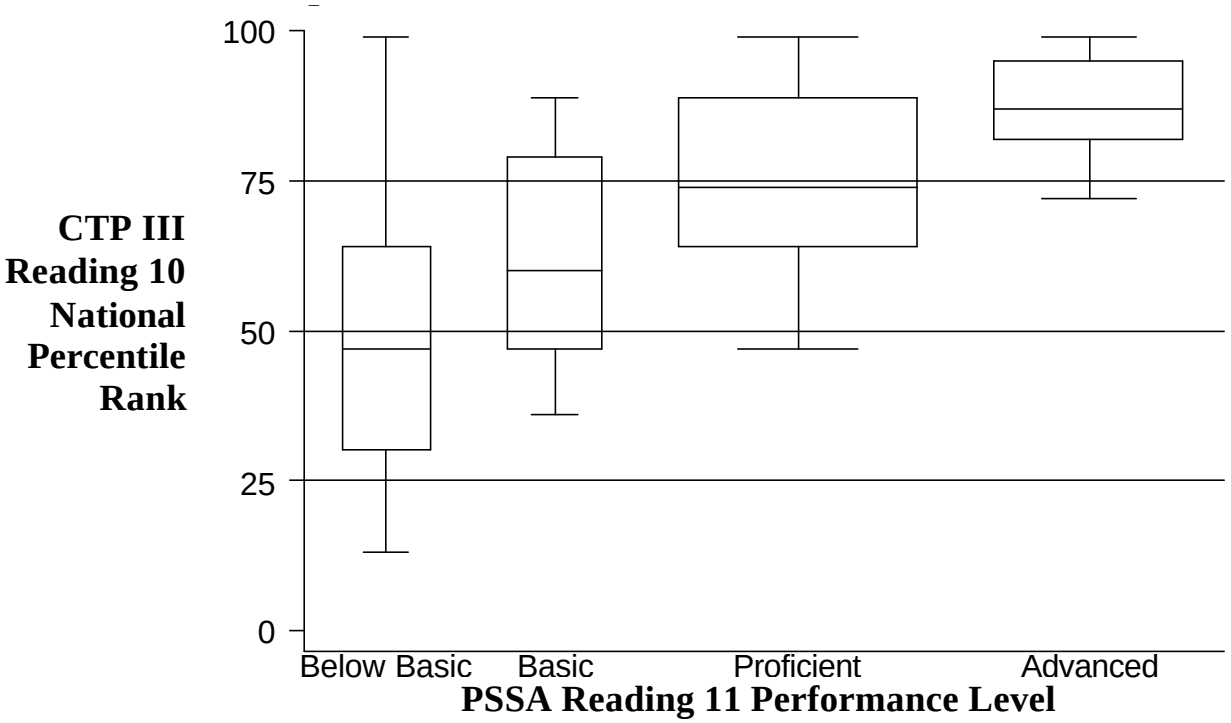


Table 19

District D		ERB-CTP III Reading 10 Percentile			PSSA Scaled Score		
		N	Mean	Std. Dev.	N	Mean	Std. Dev.
PSSA Reading 11							
	Below Basic	18	46.4	23.0	20	806	421.9
	Basic	19	61.3	17.7	22	1215	49.2
	Proficient	48	74.9	14.8	54	1419	62.1
	Advanced	38	88.0	8.1	42	1619	63.7

Graph 20

Districts A, B, &C: Comparison of Individual's Scores on the SAT I Reasoning Test (Verbal) and PSSA Reading 11

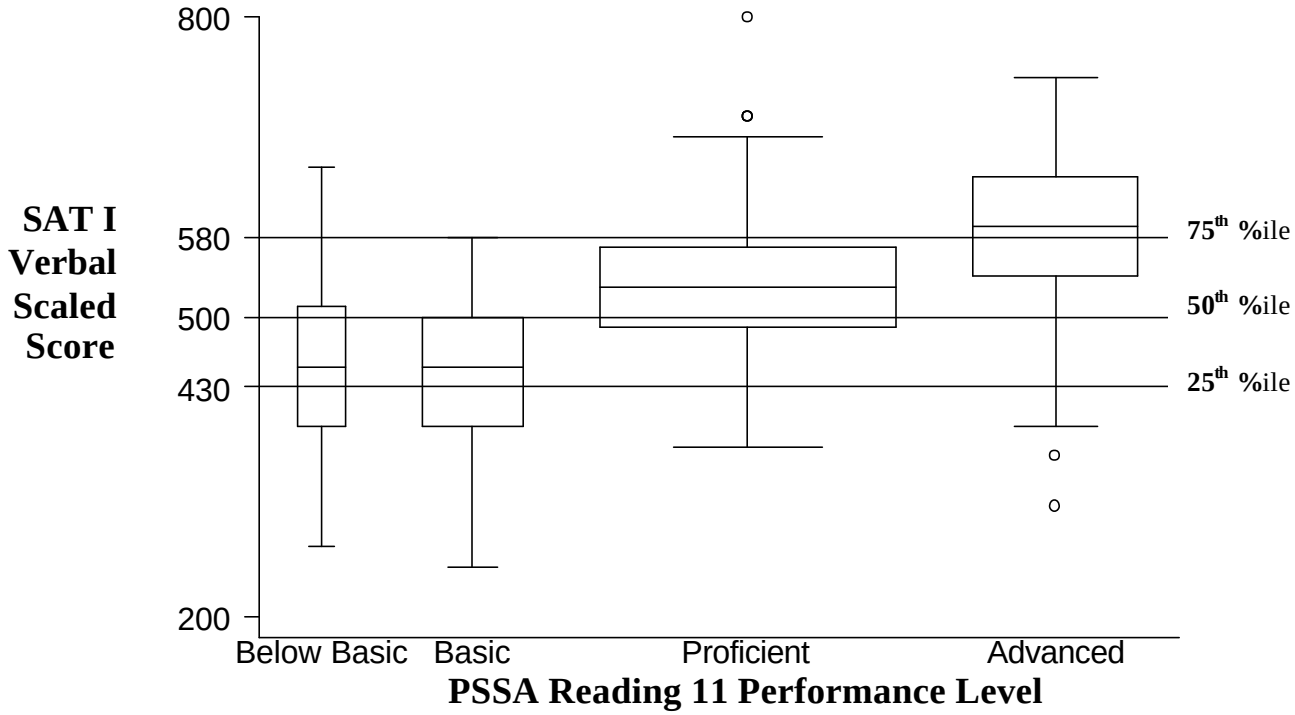


Table 20

Districts A, B, & C

	SAT I Verbal Scaled Score		
	N	**Mean**	**Std. Dev.**
PSSA Reading 11			
Below Basic	**33**	**457**	**89.5**
Basic	**67**	**446**	**73.9**
Proficient	**197**	**531**	**68.0**
Advanced	**110**	**586**	**81.8**

III. Recommendations and Final Conclusions

There can be little doubt that statewide student testing will assume an increasingly prominent role as a mechanism for providing educational accountability. While much difficult and highly skilled work has gone into the construction of the performance standards for the PSSA testing system, the validity and acceptance of them require a greater degree of disclosure and transparency in the policy-laden decisions being made.

This analysis of the recently adopted cut scores for the PSSA 5th, 8th, and 11th grade Mathematics and Reading tests should, at a minimum, give policymakers reason to commission a more comprehensive evaluation of the cut score decisions they have made.

In that spirit, we offer these recommendations:

- The State Board of Education should commission a broad review of the development of PSSA performance standards/cut scores and their use as a central part of Pennsylvania's system for ensuring educational accountability.
- The review should include, but not be limited to, a technical evaluation of the performance standard-setting process and outcomes. It should also provide for a comprehensive study of the external validity of the performance levels and associated cut scores.
- Finally, the review should consider the appropriateness of the multiple uses to which these performance levels are being put.
- The review panel should include testing experts, but also should include measurement experts from other social science fields (e.g., econometricians and demographers). It would be preferable that none of the experts have any prior ties to the PDE or any of the Commonwealth's educational associations. The panel should also include representatives of other stakeholder groups. Their final report and supporting documentation should be posted on the PDE website.
- All technical reports and documentation relating to the development of the PSSA performance standards and associated cut scores should be posted on the PDE website.
- The State Board of Education should, at a minimum, end the "multiple hurdle" approach to awarding diploma seals. PSEA considers an approach that incorporates multiple forms of assessment to be more valid and of much greater pedagogical value.

Appendix A

June 27, 2001

Mr. Michael Kozup
Director
Bureau of Curriculum and Academic Services
333 Market Street
Harrisburg, PA 17126-0333

Dear Mr. Kozup,

In accordance with the directions given to me by Dr. Lenny Lock, I make the following request for information on behalf of the Pennsylvania State Education Association. PSEA requests that the Department of Education provide:

- All technical reports or summaries created by the PDE or its contractor California Test Bureau [CTB] (redacted to remove individual students results) regarding the validation, reliability, and precision of the PSSA performance level descriptors and associated performance level scores. Please provide any information you have indicating the content validity of the tests. Are any of the items used in other assessment instruments? If so, please specify which instruments and their purpose. Has the PDE or its contractor studied the relationship between performance on the PSSA tests and any national norm or criterion-referenced tests? If so, please provide any information you have regarding the results of that analysis.

- A step-by-step summary of the statistical analyses and results used to develop the performance level descriptors and performance level scores. Specifically, we are interested in measurements of the inter-rater reliabilities for the bookmark procedure, the correlations between teacher ratings of students and their PSSA scores, as well as the precision of the bookmark and borderline groups methods. For example, when using the borderline groups method, what was the correlation between the teachers' evaluations of students and their actual PSSA scores? What were the correlations between the cut scores established through the bookmark and borderline groups techniques? What are the confidence intervals around the cut scores?

- Please provide all information regarding the criteria used by teachers in evaluating students as part of the borderline groups procedure (e.g., methodology, reliability, validity, etc.)

- In "the final confirmatory step before consideration by the State Board", the PDE reconvened 30 teachers and surveyed them to validate performance expectations and associated scores. Please provide all documentation, survey questions, results and some indication of how the 30 teachers were selected for this task.

- Any documentation relating to the external technical advice or peer review to which the performance descriptor and level setting process has been subjected. This would include any advice or review regarding the validity, reliability, and precision of the performance level descriptors and associated cut scores actually adopted by the State Board of Education.

Thank you in advance for your consideration.

Yours sincerely,

Harris L. Zwerling, J.D, Ph.D.
Assistant Director of Research

Cc: Charles Zogby
Lee Plempel
Leonard Lock
James Masters

Appendix B

Grade 5 Mathematics: Bookmark Versus Borderline Groups Performance Level Classifications
Borderline Groups Performance Level

Bookmark Performance Level	Clearly Advanced	Borderline Advanced/ Proficient	Clearly Proficient	Borderline Proficient/ Basic	Clearly Basic	Borderline Basic/Below Basic	Clearly Below Basic	Row Total	As Calculated on Sheet
Below Basic	0	6	25	48	51	**116**	**156**	402	515
Basic	2	41	178	**196**	**244**	**221**	137	1019	1226
Proficient	56	**242**	**560**	**311**	198	94	30	1491	1714
Advanced	**267**	**355**	259	70	21	9	3	984	1160
Column Total	325	644	1022	625	514	440	326	3896	4615

Grade 8 Mathematics: Bookmark Versus Borderline Groups Performance Level Classifications
Borderline Groups Performance Level

Bookmark Performance Level	Clearly Advanced	Borderline Advanced/ Proficient	Clearly Proficient	Borderline Proficient/ Basic	Clearly Basic	Borderline Basic/Below Basic	Clearly Below Basic	Row Total	As Calculated on Sheet
Below Basic	0	3	39	69	117	**101**	**133**	462	619
Basic	4	29	158	**140**	**125**	**69**	34	559	729
Proficient	45	**156**	**301**	**97**	49	21	16	685	838
Advanced	**142**	**151**	130	23	5	1		452	519
Column Total	191	339	628	329	296	192	183	2158	2705

Source: CTB/McGraw-Hill, 2001b, p. J-11

Grade 11 Mathematics: Bookmark Versus Borderline Groups Performance Level Classifications
Borderline Groups Performance Level

Bookmark Performance Level	Clearly Advanced	Borderline Advanced/ Proficient	Clearly Proficient	Borderline Proficient/ Basic	Clearly Basic	Borderline Basic/Below Basic	Clearly Below Basic	Row Total	As Calculated on Sheet
Below Basic	4	14	87	87	123	**90**	**51**	456	730
Basic	5	33	143	**108**	**91**	**28**	15	423	538
Proficient	32	**126**	**250**	**110**	70	14	5	607	709
Advanced	**121**	**154**	112	26	9	1	0	423	470
Column Total	162	327	592	331	293	133	71	1909	2447

Appendix C

Grade 5 Reading: Bookmark Versus Borderline Groups Performance Level Classifications
Borderline Groups Performance Level (Teacher Classification)

Bookmark Performance Level	Clearly Advanced	Borderline Advanced/ Proficient	Clearly Proficient	Borderline Proficient/ Basic	Clearly Basic	Borderline Basic/Below Basic	Clearly Below Basic	Row Total
Below Basic			1	8	16	**26**	**51**	102
Basic	3	7	27	**49**	**84**	**65**	49	284
Proficient	32	**89**	**243**	**116**	118	50	13	661
Advanced	**118**	**151**	133	23	16	6	4	451
Column Total	153	247	404	196	234	147	117	1498

Grade 8 Reading: Bookmark Versus Borderline Groups Performance Level Classifications
Borderline Groups Performance Level (Teacher Classification)

Bookmark Performance Level	Clearly Advanced	Borderline Advanced/ Proficient	Clearly Proficient	Borderline Proficient/ Basic	Clearly Basic	Borderline Basic/Below Basic	Clearly Below Basic	Row Total
Below Basic		1	13	20	18	**9**	**22**	83
Basic	4	19	74	**63**	**38**	**29**	15	242
Proficient	69	**102**	**182**	**74**	23	8	2	460
Advanced	**78**	**41**	43	3	1		1	167
Column Total	151	163	312	160	80	46	40	952

Source: CTB/McGraw-Hill, 2001a, p. I-11

Grade 11 Reading: Bookmark Versus Borderline Groups Performance Level Classifications
Borderline Groups Performance Level (Teacher Classification)

Bookmark Performance Level	Clearly Advanced	Borderline Advanced/ Proficient	Clearly Proficient	Borderline Proficient/ Basic	Clearly Basic	Borderline Basic/Below Basic	Clearly Below Basic	Row Total
Below Basic			3	1	1			5
Basic		2	27	**11**	**12**	**1**	5	58
Proficient	24	**42**	**102**	**16**	20	1		205
Advanced	**32**	**10**	9					51
Column Total	56	54	141	28	33	2	5	319

Appendix D

Source: Pennsylvania Dept. of Education

The PDE Adjustments to the Bookmark Performance Levels Decrease the Total Number of Students Achieving Proficient or Higher by 25,576 and Increase the Number of Failures by 42,134

5th Grade Mathematics

	Bookmark Performance Level			Borderline Performance Level			PDE Cut Score Math 5			PDE Cut vs. Bookmark	
Performance Level	Cut Scores	Number	Percent	Cut Scores	Number	Percent	Cut Scores	Number	Percent	Number	Percent
Advanced	1440	32,767	24.2	1458	29,912	22.1	1460	27,149	20.1		
Proficient	1262	45,813	33.9	1300	39,775	29.4	1300	42,538	31.4		
Basic	1129	32,563	24.1	1186	30,233	22.3	1170	32,625	24.1		
Below Basic		24,180	17.9		35,403	26.2		33,011	24.4		
Total		135,323	100.1		135,323	100.0		135,323	100.0		
Advanced + Proficient		78,580	58.1		69,687	51.5		69,687	51.5	**(8,893)**	**-6.6**
Below Basic		24,180	17.9		35,403	26.2		33,011	24.4	**8,831**	**6.5**

5th Grade Reading

	Bookmark Performance Level			Borderline Performance Level			PDE Cut Score Reading 5			PDE Cut vs. Bookmark	
	Cut Scores	Number	Percent	Cut Scores	Number	Percent	Cut Scores	Number	Percent	Number	Percent
Advanced	1445	33,616	24.9	1487	28,196	20.9	1480	28,196	20.9		
Proficient	1259	51,775	38.4	1313	45,230	33.6	1300	45,230	33.6		
Basic	1077	31,438	23.3	1201	25,057	18.6	1160	30,726	22.8		
Below Basic		17,961	13.3		36,307	26.9		30,638	22.7		
Total		134,790	99.9		134,790	100.0		134,790	100.0		
Advanced + Proficient		85,391	63.3		73,426	54.5		73,426	54.5	**(11,965)**	**-8.8**
Below Basic		17,961	13.3		36,307	26.9		30,638	22.7	**12,677**	**9.4**

Appendix D (Cont'd)

8th Grade Mathematics

	Bookmark Performance Level			Borderline Performance Level			PDE Cut Score Math 8			PDE Cut vs. Bookmark	
	Cut Scores	Number	Percent	Cut Scores	Number	Percent	Cut Scores	Number	Percent	Number	Percent
Advanced	1499	25,693	19.1	1483	28,016	20.8	1510	23,447	17.4		
Proficient	1337	33,959	25.2	1305	38,134	28.3	1300	44,871	33.3		
Basic	1203	29,702	22.0	1196	25,192	18.7	1180	25,038	18.6		
Below Basic		45,417	33.7		43,429	32.2		41,415	30.7		
Total		134,771	100.0		134,771	100.0		134,771	100.0		
Advanced + Proficient		59,652	44.3		66,150	49.1		68,318	50.7	**8,666**	**6.4**
Below Basic		45,417	33.7		43,429	32.2		41,415	30.7	**(4,002)**	**-3.0**

8th Grade Reading

	Bookmark Performance Level			Borderline Performance Level			PDE Cut Score Reading 8			PDE Cut vs. Bookmark	
	Cut Scores	Number	Percent	Cut Scores	Number	Percent	Cut Scores	Number	Percent	Number	Percent
Advanced	1509	22,284	16.6	1422	44,485	33.2	1490	27,550	20.5		
Proficient	1266	57,684	43.0	1253	39,675	29.6	1280	47,981	35.8		
Basic	1066	34,321	25.6	1160	17,402	13.0	1130	28,798	21.5		
Below Basic		19,830	14.8		32,557	24.3		29,790	22.2		
Total		134,119	100.0		134,119	100.1		134,119	100.0		
Advanced + Proficient		79,968	59.6		84,160	62.8		75,531	56.3	**(4,437)**	**-3.3**
Below Basic		19,830	14.8		32,557	24.3		29,790	22.2	**9,960**	**7.4**

Appendix D (Cont'd)

11th Grade Mathematics

	Bookmark Performance Level			Borderline Performance Level			PDE Cut Score Math 11			PDE Cut vs. Bookmark	
	Cut Scores	Number	Percent	Cut Scores	Number	Percent	Cut Scores	Number	Percent	Number	Percent
Advanced	1478	24,110	21.4	1465	25,852	22.9	1490	22,375	19.8		
Proficient	1315	26,835	23.8	1275	32,606	28.9	1310	28,570	25.3		
Basic	1188	23,380	20.7	1134	26,491	23.5	1180	25,556	22.6		
Below Basic		38,538	34.1		27,914	24.7		36,362	32.2		
Total		112,863	100.0		112,863	100.0		112,863	99.9		
Advanced + Proficient		50,945	45.2		58,458	51.8		50,945	45.1	**-**	**-0.1**
Below Basic		38,538	34.1		27,914	24.7		36,362	32.2	**(2,176)**	**-1.9**

11th Grade Reading

	Bookmark Performance Level			Borderline Performance Level			PDE Cut Score Reading 11			PDE Cut vs. Bookmark	
	Cut Scores	Number	Percent	Cut Scores	Number	Percent	Cut Scores	Number	Percent	Number	Percent
Advanced	1589	8,317	7.4	1493	19,868	17.7	1520	15,877	14.2		
Proficient	1243	58,806	52.5	1288	38,686	34.6	1290	42,299	37.8		
Basic	1000	33,900	30.3	1248	7,739	6.9	1140	26,003	23.2		
Below Basic		10,934	9.8		45,664	40.8		27,778	24.8		
Total		111,957	100.0		111,957	100.0		111,957	100.0		
Advanced + Proficient		67,123	59.9		58,554	52.3		58,176	52.0	**(8,947)**	**-7.9**
Below Basic		10,934	9.8		45,664	40.8		27,778	24.8	**16,844**	**15.0**

Grand Totals: PDE Cuts vs. Bookmark		
Advanced + Proficient:	**(25,576)**	**-3.4**
Below Basic:	**42,134**	**5.6**

The PDE Adjustments to the Bookmark Performance Levels Decrease the Total Number of Students Achieving Proficient or Higher by 25,576 and Increase the Number of Failures by 42,134.

References

1. Pennsylvania Bulletin. (2001). Vol. 31, No. 21, May 31, 2001.

2. Pennsylvania Department of Education (2001a). "PA State Board of Education Approves Performance Levels for PSSA". PDE Website: May 11, 2001.

3. Pennsylvania Department of Education. (2001b). "Summary of 2001 Pennsylvania System of School Assessment (PSSA) Reading, Mathematics and Writing School-By-School Scaled Scores." PDE Website: November 2001.

4. Zieky, M. J. (2001). "So Much Has Changed: How the Setting of Cutscores has Evolved Since the 1980s." in G. J. Cizek, (Ed.). Setting Performance Standards: Concepts, Methods, and Perspectives. (pp. 19-51). Mahwah, NJ: Lawrence Erlbaum Associates.

5. Commonwealth of Pennsylvania News Release. 2001a. "RIDGE ADMINISTRATION NAMES FOUR NEW MEMBERS TO ACADEMIC ADVISORY TEAM FOR HARRISBURG SCHOOL DISTRICT." Department of Education: January 4, 2001. Retrieved from web at: http://papress.state.pa.us/ctc/data/20010105.001.htm

6. Commonwealth of Pennsylvania News Release. 2001b. "SECRETARY HICKOK SENDS LETTER TO PSEA PRESIDENT PATSY TALLARICO." Department of Education: March 2, 2001. Retrieved from web at: http://papress.state.pa.us/ctc/data/20010302.0015.htm

7. Commonwealth of Pennsylvania News Release. 2002. "SCHWEIKER ADMINISTRATION RELEASES LATEST EMPOWERMENT LIST." Department of Education: January 4, 2002. Retrieved from web at: http://papress.state.pa.us/ctc/data/20020104.003.htm

8. Executive Summary of Governor Schweiker's Proposal 2002. Governor's Philadelphia School Proposal. Retrieved from web at: http://www.pdenewsroom.state.pa.us/newsroom/lib/newsroom/govplan.pdf

9. Pennsylvania Department of Education. (2001c). "Section I: For State Board of Education Consideration". Handout provided to the Study Session of the Pennsylvania State Board of Education. May 9, 2001.

10. CTB/McGraw-Hill. (2001a). Pennsylvania System of School Assessment, Standard Setting Technical Report Grades 5,8, 11 Reading, Submitted to the Pennsylvania Department of Education August 2001. (Cited as Reading Technical Report).

11. CTB/McGraw-Hill. (2001b) Pennsylvania System of School Assessment, Standard Setting Technical Report Grades 5,8, 11 Mathematics, Submitted to the Pennsylvania Department of Education September 2001. (Cited as Mathematics Technical Report).

12. Mitzel, H.C., D. M. Lewis, R.J. Patz, and D.R. Green. (2001). “The Bookmark Procedure: Psychological Perspectives.” in G. J. Cizek, (Ed.). Setting Performance Standards: Concepts, Methods, and Perspectives. (pp. 249-282). Mahwah, NJ: Lawrence Erlbaum Associates.

13. Hambleton, R. K. (2001). “Setting Performance Standards on Educational Assessments and Criteria for Evaluating the Process.” in G. J. Cizek, (Ed.). Setting Performance Standards: Concepts, Methods, and Perspectives. (pp. 89-116). Mahwah, NJ: Lawrence Erlbaum Associates.

14. Kane, M.T. (2001). “So Much remains the Same: Conception and Status of Validation in Setting Standards.” in G. J. Cizek, (Ed.). Setting Performance Standards: Concepts, Methods, and Perspectives. (pp. 53-88). Mahwah, NJ: Lawrence Erlbaum Associates.

15. Lewis, D.M. (2001). “Overview of the Standard Errors Associated with Standard Setting.” In CTB/McGraw-Hill. Pennsylvania System of School Assessment, Standard Setting Technical Report Grades 5,8, 11 Reading. (pp. C1-1 to C1-4). Submitted to the Pennsylvania Department of Education August 2001.

16. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for Educational and psychological testing. Washington, D.C.: American Psychological Association.

17. Cronbach, L. J., R. L. Linn, R. L. Brennan, and E. Haertel. (1995) “Generalizability Analysis for Educational Assessment.” Evaluation Comment. UCLA’s Center for the Study of Evaluation & The National Center for Research on Evaluation, Standards, and Student Testing. Summer 1995.

18. Rhoades, Sen. J. J. (2002) “Graduation Seals: an Opportunity to Verify Our Priorities” Remarks to the Pennsylvania State Board of Education. January 17, 2002.

19. Stata Corporation. (2001). Stata Graphics Manual Release 7. College Station, TX: Stata Press.

20. Gong, B. (1999). “Relationship Between Student Performance on the MCAS (Massachusetts Comprehensive Assessment System) and Other Tests-Collaborating District A, Grades 4 and 10.” Report prepared for the Massachusetts Department of Education. Dover, NH: National Center for the Improvement of Educational Assessment.

21. College Entrance Examination Board. (2001). 2001 College Bound Seniors: A Profile of SAT Program Test Takers.