HumRRO Report No.
FR-04-34

HumRRO

Human Resources Research Organization

PSSA Issues and Recommendations

Arthur A. Thacker

Prepared for: Central Susquehanna Intermediate Unit
P.O. Box 213
Lewisburg, PA 17837

May 2004

66 Canal Center Plaza, Suite 400 • Alexandria, VA 22314 • 703.549.3611 • fax. 703.549.9025
www.humrro.org

PSSA Issues and Recommendations

Introduction 2
Findings 2
Addressing Questions 1 and 5 2
Addressing Questions 2 and 4 4
Conclusions and Recommendations 6
Question 1 6
Question 2 6
Question 3 6
Question 4 7
Question 5 7
References 8

PSSA ISSUES AND RECOMMENDATIONS

Introduction

HumRRO has conducted a series of studies for the Pennsylvania State Board of Education related to the validity of the Pennsylvania System of School Assessment (PSSA). HumRRO produced three technical reports that relate the methodology and results of those studies. This report is a nontechnical summary of the findings and implications of that research. Specific questions regarding methodology or technical information are answered in the larger reports.

These studies were designed to respond to the following five questions, originally listed as part of the Request for Proposals (RFP) resulting in this research:

- Question 1. Does the PSSA adequately measure the academic content specified by the State Standards contained in Chapter 4?
- Question 2. Are the PSSA tests internally consistent and replicable?
- Question 3. Does the PSSA produce results that support decisions required by Chapter 4 regulations? These include a determination of whether a student has demonstrated proficiency in meeting State academic standards in reading, writing, and mathematics; the award of a State certificate of proficiency or distinction; etc.
- Question 4. Do the scores produced by PSSA correlate positively and significantly with pertinent scores produced on related tests such as Terra Nova, Stanford Achievement Test, etc.?
- Question 5. Were the methodologies used to determine performance levels (cut scores) reasonable and technically competent?

Findings

Addressing Questions 1 and 5

In order to answer the questions listed in the RFP, HumRRO conceptualized two types of studies. The first study focused on item content and difficulty mapping (Thacker & Dickinson, 2004), and it was designed to address Questions 1 and 5. From item information provided by Data Recognition Corporation (DRC), Pennsylvania's testing contractor, HumRRO constructed item content and difficulty maps linked to the State Standards for each grade level and subject. These maps allow the test content to be compared with the content standards and the cut scores in an easily understood manner. The maps use the cut scores as indicators to delineate items by difficulty. The report also includes a brief literature review on the standards-setting methods used in Pennsylvania and the manner in which those standards are applied to the PSSA scale scores.

Question 1. Does the PSSA adequately measure the academic content specified by the State Standards contained in Chapter 4?

- All PSSA test forms contain approximately the same number of items per academic standard.
- Sub-standards are so numerous that they are often represented by only 1 or 2 items per form.
- Overall item difficulty is similar by form.
- Item difficulty is not similar by type. Multiple-choice items tended to discriminate best at the lower and middle portions of the scale. Performance-task items tended to discriminate across the scale, with scores of 4 or 5 only reached by the very highest ability students.
- Content is not distributed evenly by item type. Content standards are written such that item type seems implied by the standard. For instance, all reading performance-task items carry a code for "reading, analyzing and interpreting literature." The standard is also assessed by multiple-choice items, but it seems clear that an aspect of the standard is tied to students' ability to respond to the performance-task prompts.
- Some PSSA tests discriminate best (have the smallest error estimates) within the lower performance categories (Below Basic and Basic).
- Very few students score 3, 4, or 5 points on several mathematics performance-task items, often leading to item parameters that are difficult to interpret.
- Some academic standards are not represented by the PSSA items.

Question 5. Were the methodologies used to determine performance levels (cut scores) reasonable and technically competent?

- Pennsylvania used both the Bookmark and Borderline Groups methods for establishing cut scores.
- Both standards-setting methods have been used for similar assessments in other states and have resulted in reasonable cut scores.
- Borderline Groups is one of the most reliable methods of setting cut scores.
- Bookmark procedures typically lead to low standard deviations compared to other methods.
- There are distinct advantages of using multiple methods for setting cut scores. Content experts are allowed to consider the results of standards-setting in relation to the purposes of the assessment and to decide through arbitration which method results in the most appropriate standard.
- Raising the cut scores by one quarter standard error was a policy decision designed to endorse rigorous student expectations and was not investigated empirically.

Addressing Questions 2 and 4

The second type of study resulted in two separate reports, both of which primarily address Questions 2 and 4. These were correlation studies designed to produce convergent validity coefficients relating PSSA with other established measures of student performance. The first report examined the relationships between PSSA and SAT results using a matched sample of students constituting roughly 90% of students taking the SAT (Koger, Thacker, & Dickinson, 2004). The second report examined the relationships between PSSA and results from several district-administered assessments (Thacker, Dickinson, & Koger, 2004). HumRRO also examined the PSSA technical reports constructed by DRC for 2001, 2002 and 2003 administrations as part of this task. Major findings follow:

Question 2. Are the PSSA tests internally consistent and replicable?

- DRC reports high reliability coefficients of greater than 0.9 for PSSA reading and mathematics tests.
- PSSA has high internal consistency reliability estimates due, in part, to the large number of items on each test.
- Convergent validity coefficients were about the same from year to year.

Question 4. Do the scores produced by PSSA correlate positively and significantly with pertinent scores produced on related tests such as Terra Nova, Stanford Achievement Test, etc.?

- Sufficient numbers of students were matched for these studies that all correlations were statistically significant ($p<0.05$).
- PSSA scale scores correlated positively with all measures studied including SAT, CTBS/Terra Nova, CAT-5 (California Assessment Test, version 5), NWEA (Northwest Evaluation Association) tests, and NSRE (New Standards Reference Exam).
- Convergent validity coefficients (correlations from different tests of the same or similar content) were appropriately high.
- For norm-referenced tests like CTBS/Terra Nova and CAT-5, and for NSRE (a student growth model test) mathematics convergent validity coefficients were typically around 0.8 and reading coefficients around 0.7. These correlations provide strong evidence that the constructs (mathematics and reading ability) measured by PSSA and these norm-referenced tests are very similar.
- For tests designed to be predictive of future student performance, such as the SAT and NWEA, coefficients were even higher. Mathematics coefficients were often around 0.9 and reading around 0.8. These correlations provide strong evidence that the constructs (mathematics and reading ability) measured by PSSA and these predictive tests are perhaps even more similar to each other than PSSA and the norm-referenced tests.

- When gain scores were considered (only for SAT), school-level improvement scores on the PSSA were positively correlated with gains on the SAT.
- PSSA scale scores also correlate positively with students' self-reported grade point average.

In addition to examining correlations between similar subjects on different tests, HumRRO also included non-PSSA subjects, particularly science, and off-subject (correlating reading scores with math scores) correlations. Those correlations were always positive and were typically not quite as strong as convergent validity coefficients. This provides a clear indication that academically strong students tend to score well in all subjects, irrespective of the measurement instrument.

HumRRO also examined the differential performance of student subgroups on PSSA and related assessments. These findings relate directly to the requirements of the federal No Child Left Behind act. It is important to remember that the main validity question is not whether subgroups score differently on PSSA, but whether the differences are consistent with those found on the comparison tests. Researchers made the following major findings:

- Gender difference effect sizes were small and nearly identical for PSSA and all comparison tests. Differences in scores related to gender were consistent for PSSA and comparison tests.
- Differences associated with socioeconomic status (students identified within the PSSA student demographic section) indicate that economically disadvantaged students did not score as well as their peers on PSSA or comparison tests. Effect sizes were moderate-to-high and nearly identical for PSSA and all comparison tests.
- SAT data included information related to students' socioeconomic status, including mother's and father's education level and family income. All were associated with scores on both PSSA and SAT. Mean student scores for both assessments increased as family income increased and as either parent's education level increased. Differences in scores were similar for SAT and PSSA for all socioeconomic variables.
- Mean scores for students with limited English proficiency (LEP) were lower than their non-LEP peers' scores for PSSA and all comparison tests. Differences were consistently higher for reading/language tests than for mathematics. Differences in scores related to LEP status were similar for PSSA and all comparison tests.
- Mean scores varied by student ethnicity for PSSA and all comparison assessments. Whites' and Asians' mean scores were consistently higher than Hispanics' mean scores, which were consistently higher than African-Americans' mean scores. Effect sizes between Whites and Hispanics and between Whites and African-Americans were from moderate to high for PSSA and all comparison tests. Differences in scores related to ethnicity were similar for PSSA and all comparison tests.

Conclusions and Recommendations

Question 1

PSSA items represent the academic content specified with a reasonable set of items per academic content area. Sub-standards are too numerous to be reported separately. Content is consistently represented across forms and item difficulty. Item difficulty is very different by item type. PSSA relies heavily on performance-task items to differentiate students at the upper end of the distribution.

Potential follow-up studies might examine the following issues:

- A closer examination of mathematics performance-task item rubrics and scoring,
- Some consideration of adding more difficult multiple-choice items to PSSA in order to better discriminate across all cut scores,
- A consequential validity study on the impact of performance-task items on classroom instruction, and
- An impact study investigating the possible repercussions of leaving some academic standards untested on the PSSA.

Question 2

Internal consistency reliability statistics are very high. Test score distributions are very similar among cohorts of students from administration to administration. Correlations among PSSA and comparison assessments are nearly identical from one year to the next.

Follow-up studies include a determination of the classification accuracy of student performance designations and potentially a similar study related to school classification accuracy. Also, impact studies could also help determine what effect being assigned to a particular category has on students.

Follow-up studies should also include monitoring the reliability measures produced in DRC's technical reports and consideration of additional measures of reliability.

Question 3

The state board chose not to have HumRRO address Question 3 as part of the research agenda. Question 3 is essentially a policy question and can only be answered based on the preponderance of evidence about PSSA. A portion of that evidence is represented by the reports submitted as part of this project. However, in order to make sound policy decisions, those reports should be considered in concert with DRC's technical manuals and other evidence regarding the validity, reliability, and potential uses of the PSSA and PSSA reports.

Question 4

PSSA scores correlate positively and significantly with all the comparison tests included in this study, including SAT, CTBS/Terra Nova, CAT-5, NWEA, and NSRE. Correlations were typically from 0.7 to 0.9 for reading and mathematics. These correlations are very strong compared to similar studies comparing state accountability tests with other measures performed in Massachusetts and Kentucky. They demonstrate a high degree of similarity between the constructs (mathematics and reading ability) measured by PSSA and the included norm-referenced and predictive tests.

Follow-up studies include a re-examination of these relationships in future years. In addition, as the uses of PSSA evolve, these types of studies should be repeated using tests with a similar purpose.

Question 5

PSSA standards setting used well-established methods. Consultants participating in the standards setting are considered experts in the field. Using two methods of standards setting allows content experts to more closely consider the standards in relation to the uses for PSSA. Raising the established cut scores by one fourth of one standard error was a policy decision and was not investigated empirically.

References

Koger, M. E., Thacker, A. A., & Dickinson, E. R. (2004). *Relationships Among the Pennsylvania System of School Assessment (PSSA) Scores, SAT Scores, and Self-Reported High School Grades for the Classes of 2002 and 2003* (HumRRO Report No. FR-04-26). Alexandria, VA: Human Resources Research Organization.

Thacker, A. A. & Dickinson, E. R. (2004). *Item Content and Difficulty Mapping by Form and Item Type for the 2001-03 Pennsylvania System of School Assessment (PSSA)* (HumRRO Report No. 04-12). Alexandria, VA: Human Resources Research Organization.

Thacker, A. A., Dickinson, E. R., & Koger, M. E. (2004). *Relationships Among the Pennsylvania System of School Assessment (PSSA) and Other Commonly Administered Assessments* (HumRRO Report No. FR-04-33). Alexandria, VA: Human Resources Research Organization.