KEYSTONE RESEARCH CENTER

Briefing Paper

412 N. Third Street * Harrisburg, PA 17101 * 717-255-7181 * www.keystoneresearch.org

PENNSYLVANIA'S NEW TEACHER TESTS - AN ASSESSMENT

Featuring An Evaluation of the Pennsylvania Professional Development Assistance Program (PDAP) Assessments by Nationally Recognized Testing Expert Barbara Plake

Summary

Beginning this fiscal year, and in each of the next four years, Pennsylvania plans to spend $4 million per year to implement a new system of "teacher tests" — the Professional Development Assistance Program (PDAP) Assessments. Each year, 20 percent of teachers are scheduled to take the assessments.[1] These tests will be used to identify schools or school districts (not individual teachers) which need professional development.

This briefing paper considers whether the new teacher tests are likely to improve teacher effectiveness or student achievement. It concludes that they are not. It therefore recommends ending the PDAP program. Some of the resources saved should be shifted to a Teacher Effectiveness Initiative (TEI) that would include assessment and professional development tied more closely to making teachers more effective in the classroom.

Our analysis of the PDAP program consists of two sections. In the first section, nationally recognized testing specialist Barbara Plake evaluates the PDAP assessments. In the second section, we briefly describe the Teacher Effectiveness Initiative and evaluate it against the PDAP assessments.

The PDAP assessments require all teachers to take a reading test geared to the reading skills that students must master by grades 5, 8, or 11. Teachers in elementary school, or in math or science at the junior high or secondary level, must also take a mathematics test. This means, Professor Plake finds, that the PDAP assessments

- assess the general mathematics and reading literacy of Pennsylvania teachers;
- do not measure how well teachers teach and do not, in most cases, measure teachers' command of the subject and grade level that they teach;
- do not measure, in lay person's terms, the right thing – how well teachers do their job – and therefore do not have what test experts term "content validity."

Plake points to other problems with the PDAP assessment program.

- The scores on one third of the mathematics topics on the PDAP assessments will be assessed by less than five test questions, making the scores in these areas unreliable.

- The high quality of a test is ordinarily documented with evidence showing that scores on the test correlate with scores on other assessment instruments. No such evidence has been published regarding the PDAP assessments. For this reason, the PDAP assessments do not have what is sometimes termed "statistical validity."

- There is no objective standard for defining when teachers' PDAP test scores require them to receive additional professional development.

- There is no objective standard defined regarding the fraction of teachers in a school or school district who must score low on the PDAP assessments before teachers are required to receive professional development.

- Lacking objective standards, the temptation will be to evaluate teachers based on performance compared to their peers, and to point to places where high proportions of teachers score in the bottom 25 percent. But since scoring in the bottom 25 percent – or the top 25 percent –- tells us little about how effective teachers are, such invidious comparisons serve no positive purpose. They could simply become another opportunity to criticize teachers and public education.

Following the Plake evaluation, the last part of the paper proposes that Pennsylvania abandon the PDAP in favor of a Teacher Effectiveness Initiative. This initiative would emphasize assessment of teachers' effectiveness in the classroom and professional development geared to directly improving classroom teaching. It should include demonstration projects that expand

- opportunities for student teachers to gain classroom experience,

- mentoring for new teachers by experienced, master teachers, and

- team teaching and other peer collaboration among experienced teachers.

In Pennsylvania, a TEI could signal a new direction for education reform. It would recognize and nurture teachers' commitment to children and to teaching well. By working with, not against, teachers Pennsylvania can achieve a revolution in teaching practice that unlocks unprecedented learning.

An Evaluation of the Pennsylvania Professional Development Assistance Program (PDAP) Assessments

Barbara S. Plake

Introduction

This report evaluates Pennsylvania's new teacher tests, officially called the Professional Development Assistance Program (PDAP) assessments, from a test design perspective.

According to House Bill 996, the purpose of the PDAP assessments is to measure teachers' command of the subject they teach – more precisely, their knowledge of the academic standards that define what students in their area of assignment must learn. If a substantial number of teachers in a school, school district, or other "school entity" score too low on the PDAP assessments – "at a level which requires additional academic opportunities" in the language of House Bill 996 — professional development programs may be implemented.

This report asks a series of common sense questions about the PDAP assessments.

- What do the PDAP assessments measure?
- Do they measure how well teachers' teach or teachers' mastery of the subject they teach?
- Will the results be reliable?
- Does Pennsylvania have well-defined standards for determining when teachers need additional professional development?

As well as the author's expertise in test design and implementation, the paper draws on interviews with a representative of the Teaching and Learning Division of the Educational Testing Service (ETS) that developed the PDAP assessments.

Barbara S. Plake, Ph.D. is W.C. Meierhenry Distinguished Professor of Educational Psychology, Director of the Oscar and Luella Buros Center for Testing, and Director of the Buros Institute of Mental Measurements at the University of Nebraska-Lincoln (UNL). Dr. Plake joined the UNL faculty in 1978 after receiving her Ph.D. in Educational Statistics and Measurement from the University of Iowa and working as a Professional Associate for American College Testing Programs. She has served the measurement community in several roles: by co-founding the scholarly journal *Applied Measurement in Education*, serving on the Board of Directors of the National Council on Measurement in Education (NCME), and serving as President of NCME in 1992-93. She has authored over 100 refereed publications and serves in an advisory capacity to many educational agencies and professional associations. Her expertise is primarily in the areas of teacher assessment literacy, state assessment and accountability, computerized testing, including adaptive testing methods, and licensure/certification testing, including setting of performance standards or cutscores. She has served as a consultant to the Nebraska, Massachusetts, Virginia, and Connecticut Departments of Education.

Pennsylvania's Teacher Tests

There are six different PDAP assessment tests total - one reading and one mathematics test for each of the three grade levels (5, 8, and 11) at which Pennsylvania defines academic "standards" that specify what students must learn. (Multiple versions of each PDAP Assessment grade and subject test will be generated by reordering the same questions, not with different questions.)

Teachers in kindergarten through grade 5 take the reading and mathematics tests geared to the grade 5 standards. Teachers in grade 6-8 take the reading tests geared to the grade 8 standards, with math and science teachers also taking the mathematics PDAP assessment. Teachers in grades 9-12 take the reading tests geared to the grade 11 standards, with math and science teachers also taking the mathematics PDAP assessment.

In mathematics, the PDAP assessments include questions on all the standards defining what students must know. Each mathematics test comprises 40 multiple-choice questions and covers seven content areas: a) Numbers, Number Systems, and Number Relationships, b) Computation and Estimation, c) Geometry and Trigonometry, d) Measurement and Estimation, e) Statistics, Data Analysis, and Probability, f) Algebra and Functions, and g) Reasoning and Connections, Concepts of Calculus. For each grade level, there are between two and eight test questions for each of these content components.

In reading, some of the standards defining what students must master are not tested on the PDAP assessments because they could not be assessed using a web-based multiple-choice format. These competencies would have required essay questions which are more expensive to grade. Some reading standards were also omitted because they used vocabulary that teachers outside of English would not be expected to know.

Each of the grade level tests in reading consist of seven reading passages (averaging 350-400 words), which are selected to be relevant to the grade level of the test. These passages cover a variety of fields, including social studies, science, language arts, narrative, and current events and issues. Each reading test has 37 multiple-choice questions that cover a) Reading Comprehension, b) Reading to Develop an Integrated Understanding, c) Reading Critically in Content Areas, and d) Reading Literature. At each grade level, there are between five and 13 items measuring each of these four areas.

Three additional guidelines were used in test development: a) vocabulary was chosen so that the questions measure the teacher's reading comprehension and not their range of vocabulary; b) specific literary terms were avoided; and c) the difficulty of the reading passages and questions does not vary markedly across the three grade levels.

For both the reading and mathematics tests, test questions were written and revised by ETS test developers in consultation with Pennsylvania teachers. No information is publicly available about who these teachers were or how they were selected. ETS then screened for questions that might be unfair to one or more subgroups due to sensitive situations, language, or other inappropriate features. These tests were then reviewed by representatives from the Pennsylvania Department of Education, classroom teachers, and college professors. These panels revised the draft tests and verified the match of the tests to the standards. These tests were subjected to two tryout sessions. The first mini-pilot was conducted to verify that the 60-minute time limits

were appropriate. A second pilot was conducted in the Neshaminy School District October 8 and 15, 2001. Information gathered at these pilot administrations provided information on test accuracy, administration feasibility, and interface difficulties. In addition, participants in the Neshaminy pilot test filled out surveys that ask questions related to the testing experience, including their perceptions of the difficulty of the tests, the ease of use of the software, and the relation of the tests to the standards.

It appears that ETS followed standard procedures in the test development process. As is ETS's practice, these tests were developed in accordance with the ETS Standards for Quality and Fairness. These standards are consistent with the *Standards for Educational and Psychological Testing*.[2] Following these standard practices does not mean the resulting test is reliable or a good assessment tool for teachers.

Do the PDAP Assessments Measure How Well People Teach?

The PDAP assessments do not attempt to measure how well teachers teach, either through direct observation or through analysis of whether student test scores improve more if their teachers scored higher on the PDAP assessments.

Instead, House Bill 996, Section 1203-A stipulates that teachers are to be tested in their knowledge of the standards appropriate for their assignment or certification. As this law has been operationalized, the relationship between the test that teachers must take (which are specified in Section 1203-B of the law) and the subject they teach is weak. For example, 5th grade standards are not applicable to a kindergarten through 4th grade teacher's level of assignment. While mathematics teachers at the middle and secondary level are not assigned to teach reading, they are required to take reading tests. So are teachers at the middle and secondary school level who teach music, home economics, and physical education.

In sum, in many cases PDAP assessment scores will not be a good indication of teachers mastery of their subject or of how well they can teach material in their discipline.

The Reliability of Component Scores

Another important test quality consideration is the reliability of the scores, which depends in part on the number of questions asked on each "competence" or skill for which results will be reported separately. According to ETS, separate scores will be reported for each of the major content components for reading and mathematics. Several of these content components, especially in mathematics, have small numbers of items for score reporting (across the three grade levels, seven of a total of 21 mathematics components have five or fewer items). In reading, only the Grade 5 test has a component with five or fewer items (Reading Literature: Literary Elements and Devices). The rest of the components, across the three grade levels, all have at least seven items and the majority have nine or more.

When there are less than five items used to measure performance, the reliability – and thus utility — of the score will likely be low in that content area.

Technical Problems with the Web-based Testing

The tests were administered via a web-based interface during a six-week window (November 1 to December

15, 2001). Although group and proctored administrations were encouraged, there was flexibility about when within this window teachers took the test. It appeared that ETS had prepared a well-designed administrative interface, with safeguards for administration problems and procedures to recover test information or restart if technical problems occur. However, there were substantial problems with this web-based delivery system when the program went operational November 1, 2001. These problems, for the most part, were related to the capacity of the web-based system to handle the volume of teachers who simultaneously attempted to take the test. When the system was designed, the expectation was that teachers would register to take the test on a more even basis throughout the six-week window. Some districts decided to have their teachers take the test during in-service days or other special days, such as election day, increasing the number of teachers attempting to sign onto the system during a limited period of time.

These administration problems created some concerns about the integrity and usefulness of the teachers' scores. Some teachers were unable to take the test and for some of these teachers, the requirement for taking the test was waived. Other teachers, who were able to start the test, experienced a complete system failure during the test. It is likely these teachers' scores will not be included in the final data set. In some cases, teachers were able to continue taking the test, but with a very slow response time that inhibited their ability to complete the tests. Some decision will need to be made about how to treat the scores from teachers who did not see all the test questions in the time allotted due to slowness of the system. If their scores are included in the final data set, summary results will be misleading as they will inaccurately report lower performance by these teachers who did not have a full opportunity to complete the test.

In addition, frustration and anxiety that may be the result of working with a slow, or non-functional, web-based delivery system may interfere with some teachers' ability to perform well on the test.

Another major concern pertains to the security of the test. This is especially a concern because there is only one version of each of the grade-level reading and mathematics tests. If teachers taking the test early disclose any content information about the tests, the use of these test scores for identifying professional development needs may be corrupted.

The Use of Test Results

Section 1207A of House Bill 996 specifies that, "beginning in the 2002-2003 school year, a school entity which determines that a substantial number of its teachers who participated in the assessment and scored at a level which requires additional academic opportunities shall, upon request, receive assistance from the Department in implementing a professional development program that is designed to strengthen the skills covered by the assessment" (p. 16). It is not clear from the legislation what constitutes a "substantial number" nor what score is "at a level which requires additional academic opportunities." A non-arbitrary approach would need to be implemented to identify what is meant by a "substantial number" of teachers to warrant professional development opportunities. Otherwise, decisions about which schools and districts are eligible for professional development may be based not on what the teachers know and are able to do, but rather on the performance of a school in relation to the overall performance by teachers in the state.

If all teachers are high in their level of performance, those scoring less well would receive professional development when they do not need it. Alternatively, many teachers who could benefit from professional development may not be provided such programs if they performed above a high proportion of other teachers. For these programs to be provided when they would actually help teachers, decisions regarding eligibility for

professional development should be based on what the teachers know and are able to do (i.e., "criterion-based") rather than on their relative performance.

Conclusion

This report evaluates the PDAP assessment program based on the information available at this time.

- The articulated intent of the PDAP assessments is to identify areas where teachers need professional development to make them better teachers. For the most part, however, the tests do not measure teachers' command of what they actually teach. Nor do they measure how well teachers actually teach. For both reasons, the PDAP assessments do not have what is termed "content validity."

- Instead, the testing program as operationalized appears to be more of an assessment of general reading and mathematics literacy for teachers in Pennsylvania.

- The match of the PDAP assessments to the Pennsylvania standards that specify what students must master by grades 5, 8, and 11 is weaker for reading than for mathematics. This means that the scores from the reading test will be a less good reflection of teachers' knowledge than will be the scores from the math test.

- Teachers' knowledge of some areas of reading and especially math will be assessed with a small number of test questions, making their resulting score on these areas unreliable.

- There is no evidence that performance on the PDAP assessments is related to how well people teach in their content area. For most high-quality tests, documentation exists showing that scores on the test correlate with other scores derived from other assessment instruments. In the case of the PDAP assessments, no published evidence exists regarding how performance relates to performance on other tests. This means that the PDAP assessments do not have what is sometimes termed "statistical validity."

- The web-based testing system proved unable to handle the volume of teachers attempting to take the test. This caused the system to reject some attempts to access the test and to respond slowly so that other teachers, in effect, had less time to complete their answers. As well as heightening teacher frustration and anxiety, these delivery problems could affect the integrity of the scores.

- Because there is only one set of test questions for each of the content and grade level combinations, and the test administration window was wide, there is a concern for security breeches that might corrupt the test scores.

- No objective procedure now exists for identifying the score level that identifies the need for additional academic opportunities. It is important that these levels be based on what teachers know, rather than on their relative test performance.

- Nor does an objective procedure exist regarding what is a "substantial number" of teachers who need additional academic opportunities through professional development workshops. Without clarification of what constitutes a substantial number, it is possible that decisions about eligibility for professional development programs will not be made in a fair and equitable manner.

A Teacher Effectiveness Initiative: An Alternative to the PDAP Assessment

Stephen Herzenberg

The policy issue central to the debate about Pennsylvania's new teacher tests is not whether teacher assessment and professional development can improve educational outcomes. The issue is what kind of assessment and professional development would most improve such outcomes.

In prior research, Keystone Research Center has recommended an approach to professional development quite different from the PDAP program — a "Teacher Effectiveness Initiative."[3] This proposal is motivated by research showing that teacher effectiveness improves when teachers receive more mentoring early in their career and have opportunities to reflect together with other teachers regarding classroom practice. Expanding such opportunities can help overcome the traditional isolation of American teachers in separate classrooms. This isolation reinforces a tendency to teach in customary ways – the same way this year as last year – and to lecture to passive students without evaluating whether students are actually learning.

The TEI Keystone has proposed would provide demonstration grants to expand teachers' opportunities to receive (or provide) mentoring and to collaborate with their peers. Demonstration grants could be provided

- to innovative teaching training programs that seek to overcome the disconnect of teacher education programs from the classroom;

- so that new teachers during their first two years in the profession receive extensive mentoring from "master teachers" (selected both because they are good teachers and because they are good mentors);

- to innovative proposals for peer collaboration that include evaluation of whether the innovative approaches raise student achievement.

In addition to beginning a modest demonstration program, we recommend that the Governor commission a feasibility study, with extensive input from practicing teachers and their professional associations, and make recommendations to his successor regarding a more comprehensive program to improve teacher effectiveness. Connecticut has demonstrated that a sustained bipartisan initiative to expand professional development linked tightly to classroom practice can move a state to the top of the state achievement rankings – and near the top of the international rankings (see Box 1 below).

Table 1 compares our proposed TEI with the PDAP Assessment based on eight principles of effective professional development distilled from the research on professional development by Professor Ulrich Reitzug of the University of North Carolina.[4] Table 1 shows that the PDAP Assessment program fails to meet all of the research–based characteristics of an effective professional development program. A well-designed TEI, by contrast, would meet these criteria. More important, it could improve student achievement.

Pennsylvania's new teacher tests emerged out of a Ridge Administration education policy that placed a higher priority on criticizing teachers and public schools than on improving educational outcomes. Governor Schweiker could signal a new direction in education policy by asking the legislature to cancel the PDAP assessments and launch a professional development initiative that would raise student achievement.

Table 1. Evaluating Professional Development Alternatives: Pennsylvania's Teacher Tests vs. A Teacher Effectiveness Initiative

Characteristics of Effective Professional Development Assessment and Training	*Does Each Alternative Have Recommended Characteristics?*	
	PA's Teacher Tests and PDAP Program	*A Teacher Effectiveness Initiative*
Decisions about professional development should be made within schools rather than at higher levels (solely top-down planning alienates teachers)	No – decisions about when to deliver professional development determined by performance on the teacher tests, not by teacher or school-level perceptions of need for professional development	Depends on TEI implementation
Professional development must be focused on instruction and learning	Unlikely – assessment tests basic skills therefore it is likely that training will also address basic skills	Yes – classroom observation, joint lesson planning, and evaluation of new instructional approaches built in
Professional development activities must take place over an extended period of time	Probably not – PDAP envisions testing of each group of teachers once in five years, with professional development keyed to performance on the tests	Depends on design of the program
Professional development activities should model effective pedagogy	Unlikely –since assessment tests basic skills, professional development likely to address basic skills not model effective pedagogy	Likely to include opportunities for observation of effective teachers
Professional development workshops must be supported by modeling and coaching for teachers once they return to the classroom	No provision for modeling and coaching in the statute or current program planning	Yes – a TEI program could be designed to satisfy this criterion
Professional development should focus on communities of practice (i.e., groups of teachers that teach the same subjects and age groups) rather than on individual teachers	Teacher testing and professional development will be delivered at the group level. But PD likely to focus on weaknesses identified by the test not on what groups of teachers who teach the same subject see as most likely to help them learn from each other	Yes – a TEI program could be designed to satisfy this criterion
Effective professional development requires that continuous inquiry – constantly asking what teachers are trying to accomplish and whether it is working – be embedded in the daily life of the school	No – PDAP will be divorced from the daily life of the school	Yes – a TEI provides an opportunity to embed continuous inquiry into the daily life of more schools
Principals and other school leaders must provide proactive support for professional development initiatives	Legislation is unclear regarding the involvement of principals and school leaders – initiative comes from the state	Yes – principals and school leaders would have key roles in crafting TEI initiatives

Note: Criteria for evaluating professional development are from Ulrich C. Reitzug, "Professional Development," in Alex Molnar (ed.), *What We Know About Effective Public Schools* [tentative title] (Phoenix: Education Policy Studies Laboratory (EPSL), Arizona State University, forthcoming).

Box 1. Connecticut's Long-term Effort to Improve Teaching and Learning

Since 1986, the state of Connecticut has demonstrated the value of professional development that is closely tied to improving classroom practice.[5] The historical roots of this effort go all the way back to the 1974 to 1983 period when a visionary state Commissioner of Education, Mark Shedd, reshaped the state Department of Education into a proactive learning organization staffed by graduates of leading research universities. Shedd also recognized teachers as a critical lever for education reform and targeted four issues critical to teacher quality: recruitment, initial preparation, induction, and on-going professional development.

The initial focus of Connecticut's efforts was on beginning teachers. The state recognized that measuring knowledge on standardized tests was not a sufficient indicator of teacher quality. It did, however, assess prospective teachers' knowledge in their content area via 23 different subject matter tests (as opposed to tests just in reading and math as in the PDAP assessment). These tests were combined with requiring field experiences before teachers could be certified in any subject domain and grade level.

New teachers each received a school-based mentor or mentor team for their first year through the Beginning Educator Support and Training (BEST) program. Mentors received 30 hours in professional development to help them become better at supporting new teachers. Today, second-year teachers must present a subject-specific "portfolio" of their work for assessment, which includes videotapes of two featured lessons, and an example of how they evaluate student learning. Beginning teachers who score below level two (basic) on a scale from zero to five are eligible for the third year in the BEST program. Over time, Connecticut's professional development and assessment programs have becoming increasingly subject-specific. For example, the state dropped a generic classroom observation instrument on the grounds that what it measured was covered by the more subject-specific portfolio assessment.

By one estimate, 40 percent of Connecticut's teachers have served as assessees, assessors, mentors or cooperating teachers under either the initial beginning teacher performance assessment or the newer portfolio approach. By the year 2010, 80 percent of all the state's elementary teachers, and nearly as many secondary teachers, will have participated in new subject-matter specific portfolio assessment in some role. In surveys, 80 percent of assessors say this role has improved their own teaching. Two-thirds of mentors say that mentoring has improved their own teaching.

Simultaneous with professional development reforms, Connecticut sharply increased teacher salaries. The combination of economic gains and putting teachers at the center of educational improvement have made Connecticut a more attractive place for teachers. The resulting increase in teacher supply has enabled Connecticut to raise entry-level standards. Thus, paradoxically, Connecticut's approach to professional development has almost certainly resulted in teachers who would score very well on the PDAP assessments of general reading and mathematics literacy.

By 1998, Connecticut's ranked first in the nation in reading and mathematics on the National Assessment of Educational Progress in fourth grade. In eighth grade, it had the highest share of students scoring at or above proficient in reading and was the only state to perform significantly better than the U.S. average in writing. A 1998 study found that, in the world, only Singapore would outscore Connecticut students in science. The more than 25 percent of Connecticut's students who are black or Hispanic substantially outperform their counterparts nationally.[6]

FOOTNOTES

[1] Tom Ridge, *Governor's Executive Budget* (Harrisburg: Commonwealth of Pennsylvania, February 6, 2001), p. E14.22-23.

[2] American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing* (Washington, D.C.: American Educational Research Association, 1999).

[3] See the education section of Stephen A. Herzenberg and Howard Wial, *Steal This Agenda: A Blueprint for a Better Pennsylvania* (Harrisburg: Keystone Research Center, 2000); for the intellectual foundation behind the idea of a Teacher Effectiveness Initiative, see Stephen A. Herzenberg, John A. Alic, and Howard Wial, *New Rules for a New Economy: Employment and Opportunity in Postindustrial America* (Ithaca: Cornell/ILR Press, 1998), chapter five and pp. 143-144. A Teacher Effectiveness Initiative is also part of the education policy recommendations of United Pennsylvanians, a coalition of leaders from the religious, education, labor, business, civic/consumer, and policy communities (see the Education Issues Paper of United Pennsylvanians, on line at www.unitedpa.org).

[4] Ulrich C. Reitzug, "Professional Development," in Alex Molnar (ed.), *What We Know About Effective Public Schools* [tentative title] (Phoenix: Education Policy Studies Laboratory (EPSL), Arizona State University, forthcoming).

[5] Suzanne M. Wilson, Linda Darling-Hammond, Barnett Berry, "Teaching Policy: Connecticut's Long Term Efforts to Improve Teaching and Learning," mimeo, March 2000.

[6] The Connecticut success story is consistent with analysis of eighth-grade math teaching in the United States and Japan. In cross-cultural comparison, the lack of opportunities U.S. teachers have for planning and collective reflection stands out in sharp relief. One of the basic "continuous improvement" methods in Japan is via "lesson planning" – teams of teachers spend substantial time during a year designing a single lesson, watching it taught, refining the design, and presenting it again before groups of teachers drawn from multiple schools. The most successful lessons are written up, and now web available, as curricular resources for other Japanese teachers. Students end up with deeper explanations of mathematical reasoning that translate, in turn, into better scores on international tests. James W. Stigler and James Hiebert, *The Teaching Gap: Best Ideas from the World's Teachers for Improving Education in the Classroom* (New York: the Free Press, 1999).