ESTABLISHING A SYSTEM TO EVALUATE ASSESSMENTS OF STUDENT OCCUPATIONAL SKILL ATTAINMENT

Paul M. Munyofu, Educational Research Associate
Pennsylvania Department of Education

ESTABLISHING A SYSTEM TO EVALUATE ASSESSMENTS OF STUDENT OCCUPATIONAL SKILL ATTAINMENT

Abstract

The state of Pennsylvania has been engaged in training students and assessing their Occupational Skills attainment for many decades (Kapes, 2001). Pennsylvania Department of Education (PDE) established a system of recognizing high achievement through a Skill Certificate Program which utilized 14 national and local tests. Initially the Commonwealth established a Pass/Fail decision process for one such test outlined in a report entitled: *An Evaluation of Pennsylvania Occupational Competency Written Exams Administered During 1975-78* (Kapes and Funk, 1978). However there was no system in place for evaluating and approving tests. The project in this document outlines the process for establishing and implementing a system for evaluating potential tests that might be used to measure student occupational skill attainment and determining job-readiness.

Introduction

For many years, Pennsylvania has been engaged in training and qualifying young people for employment in industry. As far back as 1984, Governor Richard Thornburgh proposed changes in his initiative “Turning the Tide: An Agenda for Excellence in Pennsylvania Public Education.” (Commonwealth of Pennsylvania, Department of Education, 1983) This was in response to a national report to the Nation and the Secretary of Education, United States Department of Education by the National Commission on Excellence in Education, (April 1983) that found that the middle group of students was not adequately served by the educational system.

In response to that plan, Jerry Olsen, State Director of Vocational Education, established a committee to review occupational testing available for use in vocational programs. The committee reviewed four or five testing programs in other states. Ohio’s testing program lacked security. The V-TECS had an open system item bank where teachers would select the items they wanted. It was determined that the National Occupational Competency Testing Institute (NOCTI) was the best alternative although they had only a limited number of tests available. At the cost of $250,000 the Bureau contracted NOCTI to develop a bank of items for 40 tests that coincided with the state’s program areas. The contract included administering, scoring and reporting the test results in the form of national norms for the 40 Student Occupational Competency Achievement Tests (SOCATS).

The Pennsylvania Skills Certificate Program

These tests were chosen because, among other qualities, they had three components. There was a cognitive portion of 80 multiple-choice items, a written portion with 150 to 200 multiple-choice items, and a performance hands-on portion of 3 to 10 tasks evaluated by an industry practitioner. The SOCAT tests were developed to reflect the program description as written by the National Center for Educational Statistics in a Classification of Instructional Programs (CIP). However, for the Pennsylvania Skills Certificate (PSC), Pennsylvania Department of Education program descriptions were also used to ensure that tests were relevant to the programs being offered in Pennsylvania. Because they were norm-referenced tests, individuals who performed at or above the national norm were awarded the PSC in recognition of high achievement. The seven purposes of the PSC program were to:

1. Provide a standardized method in all areas of “vocational/technical” education to determine the level of occupational competence of students who complete programs at the secondary level;
2. Provide a testing system based upon competencies that have been identified by workers in the respective occupations as essential for the entry-level worker;
3. Implement a two-part system of student assessment wherein the manipulative skills of the occupation are tested in addition to the theoretical concepts that are normally covered in competency evaluation;
4. Provide standards whereby teachers and administrators in career and technical education can compare the progress of a whole class or of an individual student with other students who have been enrolled in similar programs. Statistics are available by individual states and by the nation as a whole;

5. Assist local educators in the interpretation of test results that can be used to improve curriculum and instructional programs;
6. Establish employer recognition of the PSC standard. Since personnel from the various occupations have been involved in the identification of competencies upon which the tests were constructed, it is important that each local teacher and administrator recognize the value of the SOCAT score when exploring placement with the prospective employer. SOCAT and the PSC were designed for job placement; and
7. Support and conduct research appropriate to substantiate the validity and reliability of the SOCAT tests and career and technical training. (PDE, 1987)

When the Carl D. Perkins Vocational Education Act (PL 98-524) was passed, the Pennsylvania Council on Vocational Education (PACVE) was required to advise the State Board on policies that the board should pursue to strengthen vocational education in the state. The council made three recommendations that (a) the Board and the Department of Education expand the standardized process for measuring secondary vocational education occupational competencies as provided in the Pennsylvania Skills Certificate Program; (b) the Pennsylvania General Assembly approve the request of the governor to increase the reimbursement for adult vocational education programs offered in secondary education institutions as well as support for adult vocational-technical education through reimbursement of noncredit vocational Full-Time Equivalents at community agencies; and (c) the State Board and the Department of Education undertake an evaluation of the postsecondary technical education system to determine if there was a need for additional postsecondary technical education programs and services (Pennsylvania Council on Vocational Education, 1991).

A review of the various states' measures and standards of student occupational skill attainment indicated that every state had implemented some form of assessment. However there was no uniformity in developing measurement approaches that could be characterized as efficient or comprehensive. Some measured skill attainment using checklists that an instructor maintained; and at the end of the program the instructor would create a list of students who had demonstrated mastery of the required amount skills. Some states merely maintained an attendance record to determine whether a student had invested enough seat-time as an indicator of skill attainment. Others relied on locally developed and state approved tests that were presented as valid and reliable measures. Yet others utilized off-the-shelf tests of employability skills. Only a few states had invested efforts in choosing appropriate tests that were aligned with recognized industry standards. One state elected to develop tests based on the state programs.

Statement of the Problem

Under Perkins Act's accountability all students who completed a career and technical education program were expected to attain the knowledge and skills that meet program identified as industry validated career and technical skill standards. This attainment was measured as the percentage of career and technical education program completers who achieve a level of competency at or above the national norm on the NOCTI Job-Ready Assessment. At first Job-Ready competency was determined on the basis of students passing the state's Nurse Aide Training and Competency Evaluation

Program test, or on a set of three out of eight Automotive Service Excellence tests. Later that number was increased to 13 other tests that could be administered in lieu of the NOCTI for the purpose of awarding the Pennsylvania Skills Certificate.

The PSC was awarded through any of the following five ways (Commonwealth of Pennsylvania, Department of Education, 2007 Guide). First, by default, performance at or above the advanced level on the criterion-referenced NOCTI tests' written and performance components. All NOCTI tests are routinely benchmarked by teaches and industry representatives using the Nedelsky[1] method to establish cut scores for advanced, competent, basic levels. Second, passing state boards' or national licensure and certification tests. This included the boards of cosmetology, licensed practical nursing, and nurse aide tests. Third, passing a national industry-credentialing test such as American Welding Society (AWS), Industry Competency Exam (ICE), the American Culinary Federation (ACF), the American Hotel and Lodging Association (AHLA), and Electronic Technician Associate (ETA). Fourth, passing selected bundles of student end-of-program tests. This included the Computer Technology Industry Association (CompTIA), the National Automotive Technicians Education Foundation (NATEF), the National Institute for Automotive Service Excellence (AYES) and the corresponding Automotive Youth Educational Systems (AYES), and the National Institute for Metalworking Skills (NIMS) where four core tests were necessary to earn the certificate. Finally, one locally developed test by a national chapter was surprisingly allowed as an alternative to the NOCTI.

There was no clear standard system for accepting or rejecting some tests and not others. Some educators complained that they were being held to a much higher standard of performance than on the NOCTI. Others wanted to utilize any industry test regardless of depth or other considerations.

Pennsylvania also failed to meet its accountability obligations to report complete data on the Consolidated Annual Report (CAR) the number of program completers who performed at the Competent or Advanced levels on all PDE-approved tests. Some test developers could not provide data on the number of students tested and the number who achieved scores at the various levels, let alone the student demographics.

The purpose of the study was to determine whether the available tests met the criteria of valid and reliable tests that accurately measured student occupational skill attainment. The particular research questions were as follows:

1. Were the tests developed according to the established standards for educational and psychological assessments, possessing an adequate measure of rigor, relevance and content coverage as validated by industry representatives?
2. Did the tests validly measure both the cognitive and psychomotor components of the skills necessary for the respective industries?
3. Was there sufficient test security in place to ensure that test results accurately reflect the student's ability?
4. Did the test provider have the capacity of reporting comprehensive results, at the students, classroom, school, state and national levels in order that valid policy decisions can be made concerning the quality of career and technical education?
5. Was there an adequate alignment of test items to industry standards with respect to industry coverage, depth of knowledge and clarity of scoring rubrics?

Methodology

To address these concerns, it became necessary to launch a formal process of evaluating and approving all tests that would be utilized for the purpose of awarding the Pennsylvania Skills Certificate in recognition of student occupational competency.

The bureau reviewed the literature in order to identify major and critical characteristics of a valid and reliable assessment system. These characteristics were divided into four categories. An acceptable occupational test must conform to the Standards for Educational and Psychological Testing (1999). The test had to measure both a theoretical knowledge component and a practical, hands-on performance component. It was the bureau's expectation that a competent entry-level worker would need to draw on a compendium of tools and strategies to solve any task assigned. This test would also measure an entire spectrum of the chosen field with a sufficient depth of knowledge. The test would also possess an adequate measure of validated rigor, relevance and content coverage as validated by industry representatives.

A second category in evaluating the tests was technical issues related to test development. Was there sufficient documentation of reliability and validity studies conducted on the individual test items and on the entire test as a whole? Was there a professional review published in references such as Buros Institute's *Mental Measures Yearbook* and *Tests in Print*, or *Test Reviews Online*? These and other technical considerations would be established when indices such as reliability coefficients or point-biserial correlations are met. A complete technical manual is a necessity.

A third category was test administration and security. Many tests are housed on the web. They are routinely downloaded by the instructor. The tests are administered and scored by the instructor. Then the results are reported to the credentialing organization for an official credential or certificate. If the test can be administered online, are there any security safeguards in place to guarantee that the test's integrity is not breached? When there is a hands-on component to the test, this is often limited to a checklist of vague competencies. There is no consistent way of determining when a student has passed – and/or the level of performance the student has obtained. This weakness would be corrected with a precise scoring rubric that is applied by an industry practitioner when judging performance of a potential entry-level employee.

The fourth category of the test evaluation was the reporting capability. End-of-year summative assessments should yield enough for a school to evaluate its program, identifying strengths and weaknesses so that they may make appropriate changes and modifications for improvement. The report needs to contain data on individual students' achievement that must be reported to the teacher, the school administrator, the state and the federal government for accountability. A capable test developer should possess the capability to produce complete and reliable data, including student demographics as required by the Perkins Act.

Instruments Used

Appendix B is the questionnaire used to combine all the four categories into a single checklist. The checklist also served as an indicator for test providers who would like for the department to consider approving their tests for use in measuring skill attainment. A companion to this questionnaire was Appendix C. This was used as an initial rating of the responses to the questionnaire. The rating matrix contained weights attached to specific critical criteria. For example in the test development category a minimum score of 40 was necessary for a test to be considered for further review. This would be obtained if the four critical items 2, 3, 4 and 10 were met. In this matrix some items were critical and were identified as MUSTS. Some items considered essential could be overlooked if the test developers promised that at some later date they would document their development. These items were labeled ACCEPTABLE IF PROMISED. The third set of items was desirable but could be ignored if they were absent. They were consequently labeled TOLERABLE IF ABSENT.

Data Collection Procedures

With these criteria, the questionnaire (Appendix B) was sent NOCTI, NIMS and the National Center for Construction, Education and Research (NCCER). NOCTI tests were reviewed first because they were the default tests. In the absence of an alternative, NOCTI tests were administered. NIMS tests were evaluated because PDE had entered into an exclusive contract with NIMS to assess students in metalworking skills. NCCER was reviewed next because of the company potential to provide wide coverage in residential and commercial building trades. Each questionnaire was accompanied by an invitation to participate and a description of how the entire evaluation process would be carried out.

The questionnaire responses were evaluated by a team of 3 independent raters following a voting matrix (Appendix C). If the raters found sufficient indication that the test should be considered further a team would visit the test developers' headquarters to verify the information indicated on the questionnaire. The two-day second phase of the evaluation process would focus on the technical aspects of the test. This included an examination of the technical manuals, test blue prints, item analyses, validation processes, and a quick look at an actual test layout. At the exit interview there would be discussion about the test strengths, weaknesses, suggested improvements, and a tentative approval of the tests pending the results of the third and final review by content experts.

An alignment study through the Wisconsin Center for Educational Research's web tool (Webb, 2007) was performed by 5 to 10 subject matter experts from Pennsylvania. These instruments were used by industry representatives who typically employ Pennsylvania's graduates. A few teachers at the secondary and postsecondary levels were included in the panels. The panel was to determine (a) whether there was a match between Pennsylvania standards and the test items; (b) the depth of knowledge required to perform successfully on the test and on the job; (c) the clarity and accuracy of test items including graphics, and (d) the completeness of the scoring rubrics for the performance component of the tests.

Results

All three test providers scored high enough on the voting matrix as rated by three independent reviewers. There was consensus among the three reviewers that the scores warranted a step forward in the evaluation process. A two-member team visited the individual providers' respective headquarters to verify the information provided on the questionnaires. The team specified the types of documentation needed to support their questionnaire answers. Focusing on the technical aspects of the test development, the team examined, if available: (a) the technical manuals, (b) lists of subject matter experts who took part in test development, (c) item analyses that indicated the quality of the items, including p-values, point-biserial correlations and reliability indices, (d) revision and validation activities on individual items and total tests, (e) test blueprints that included weightings of critical competencies, (f) guidelines for the training of independent industry evaluators of the performance component of the test and the scoring rubrics for individual tasks. The team also reviewed security measures in place for online testing and contingency plans that addressed security breaches. The team examined sample statewide reports for individual students, for schools, for the state, and national comparative data. This was to include an interpretation manual at all levels. Finally the team examined actual test booklets for readability, accuracy and layout. Additional information was requested on the spot as necessary for clarification.

In one of the reviews it was discovered that there never was a technical manual for their battery of tests. The psychometrician knew that the manual would contain such information as the test development and revision processes, item-response theory analyses, industry validation activities. Yet there was no documentation that those activities had been carried out. Examples of the missing documents were as follows: a list of subject matter experts who wrote, revised or reviewed the items; a blueprint that would show the number of items necessary to adequately cover the area, or program, or cluster; and a process for determining a performance standard indicating how good is good enough for a candidate to be deemed competent.

Test security was among the weakest aspects of the tests reviewed. In one instance the performance component of the test was available on the web so that the instructor could download it, administer the test, and send the results to the testing organization for validation. In another, a student manufactured a piece according to published specifications and the instructor sent the piece to a testing center for judging. The instructor signed off that the piece was actually manufactured by the student.

None of the three tests reviewed in this round addressed any external reviews of their tests made by professional external reviewers. Nor did they attempt to compare themselves against other tests that purport to measure competency in the same area. Item 5 on the questionnaire was “Does the test blueprint compare favorably with that of a NOCTI assessment in the same area ensuring that similar areas are being covered to the same depth?” Though the other two said they were better, no documentation was offered to show that the necessary concurrent validity study had been performed. At the writing of this document the tests have not been reviewed or listed in Buros Institute’s Tests in Print, Test Reviews Online, or Mental Measurement Yearbook.

There was universal reference to national industry standards for each test. However none of the providers showed documentation that there was a real match between their tests and the standards they intended to measure. There was little indication of the desired depth of knowledge needed to demonstrate levels of performance of entry-level workers. When the department performed its own alignment study on a sampling of the tests, there was a sizable measure of content coverage. However that coverage was shallow in terms of depth of knowledge. It is understandable that such national tests might be generic enough to allow individuals of varying abilities and emphases to be able to utilize them. Nonetheless, a certain degree of discrimination needs to be a part the assessment framework. Real global competition does not take place at the bottom but rather at the cutting edge of the industry. End of program tests should contribute effectively to that fray.

Discussion

For the most part, the tests were developed according to the established standards for educational and psychological assessments. The onsite review was quite indispensable because the team saw documentation aligning the standards to the tests. Determination of rigor and content coverage were described in the technical manuals, including lists and demographic information about the subject matter experts who took part in the test development. Research question 1 was satisfied.

All three sets of tests did contain a written component and a performance component as required in the voting matrix. In all cases there was detailed description of how to evaluate the hands on tasks performed by the students. While NOCTI required the performance component be carried out in the presence of an industry representative, the other developers had their students do the tasks in presence of the instructor who certified their originality. NIMS required the manufactured object to be evaluated by an outside industry representative, adding more authenticity to the process. However having tasks housed on the web was a concern because students could practice on them indefinitely until they practice one more time in front of their instructor. In spite of this concern, research questions 2 and 3 were essentially satisfied.

NOCTI and NIMS have records of producing comprehensive end-of-year reports containing demographic information such as gender, ethnicity and special population categories required by federal regulations on student occupational skill attainment. At the same time, test results were routinely provided to instructors on each student achievement in each category. Aggregate results were then compiled at the classroom, state and national levels. This information made the tests utile in program improvement plans as well. The instructor could continue in areas of strength and channel more efforts in those areas where performance was below the state or the national levels. NIMS test results were reported with demographical information. However, school reports were limited to pass/fail declaration for each student and for each test. Because there was sufficient data necessary for decision making research question 4 was satisfied as well.

All tests appeared to provide acceptable industry coverage. One limitation was that these tests were generic, intended to serve as entry qualification in the entire industry. Yet not all career and technical education programs are designed in the same way. School administrators were selective in focusing on those aspects of industry that best serve their students and their local employer needs. Therefore, there was an expected difference between what was taught and what was tested. Still, the segments covered in the schools were well aligned to the test items, and research question 5 was well addressed.

Conclusions

The results of the study suggest that there is a real benefit in conducting a formal evaluation of tests used to measure skill attainment. A test provider's claimed assertions about their test can only be verified through an investigation such as described in this study. If a user is not able to conduct such an evaluation, then a reputable professional evaluator could perform the analyses. Many tests have been evaluated and results are periodically publicized. Depending on the intended use of assessments or the results of those assessments, the user is encouraged to determine their validity before administering the assessment instruments.

Not all tests are suitable for administration in all situations. Some career and technical education institutions might be interested in modular assessments that measure mastery in segments of their programs. For these institutions, an end-of-program assessment would not be appropriate. Some might want to focus on particular populations, such as special needs students functioning with limiting Individualized Education Programs (IEPs). For these, there would be little benefit derived from comprehensive tests. It is suggested that to get the most out of these assessment instruments the user is encouraged to evaluate them and, if necessary, customize to individual needs.

Footnotes

[1] Developed by Nedelsky in 1954, this content method sets "absolute standards" for setting a cutoff score on multiple choice examinations (Meskauskas, 1976). The method is based on the theory that marginal test takers (i.e., test takers who possess relatively low levels of the KSAs tested on an examination) will eliminate as many incorrect choices from an item and then guess from the remaining alternatives (Livingston & Zieky, 1982; Meskauskas, 1976).

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington DC.

Commonwealth of Pennsylvania, Department of Education. (2006). *A Guide to Student Occupational Competency Testing in Pennsylvania, 2006-2007*. Harrisburg, PA

Commonwealth of Pennsylvania, Department of Education. (1987). *A Manual for P.S.C. Coordinators*. Harrisburg, PA

Commonwealth of Pennsylvania, Department of Education. (1983). *Turning the Tide: An Agenda for Excellence in Pennsylvania Public Education*. Harrisburg, PA

Kapes, J. T. (2001) *Pennsylvania Pass/Fail Cutoff Scores for NOCTI Written and Performance Exams Based on 2001 National Norms*. Pennsylvania Department of Education.

Kapes, J. T. and Funk, G.W. (1978). *An Evaluation of Pennsylvania Occupational Competency Written Exams Administered During 1775-78)* Pennsylvania Department of Education.

Livingston, S. A. & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

The National Commission on Excellence in Education, 1983: *A Nation At Risk: The Imperative For Educational Reform*.

Nitko, A. J. *Using a Mental Measurement Yearbook Review and Other Materials to Evaluate a Test*. Buros Institute, 2005

Pennsylvania Council on Vocational Education. (1991). 22nd Annual Report. Harrisburg, PA

Webb, N. L.(2007) *Wisconsin Alignment Protocol*, Wisconsin Center for Educational Research.

Appendix A
Occupational Performance Levels Descriptors

Advanced

The Advanced level reflects mastery of competence and understanding of academic/career and technical skills and knowledge required for advanced placement in employment and/or postsecondary educational institution. Students with this score "would function like a journeyman, better than an entry level worker and should require minimal supervision on the job"

Competent

The Proficient level reflects a high degree of acquisition of academic/career and technical skills and knowledge required to enter employment and/or postsecondary education. Students with this score "would function better than an entry level worker and should require little supervision on the job"

Basic

The Basic level reflects an adequate attainment of academic/career and technical skills and knowledge required to enter employment or postsecondary education. Students with this score "would function at an entry level, but would require some assistance on the job"

Below Basic

Below Basic level reflects a partial acquisition of skills and knowledge needed to perform a given assignment/task/operation on a job. Additional instruction and/or assistance are necessary in order for the student to successfully complete specific assignments. Students with this score did not acquire the minimum skills "required for the occupation"

Appendix B: Checklist for Evaluating an Occupational Skill Test

Test : ______________________________ Industry or Professional organization: ______________________

Date of release or version:______________________ Type of test: ❐ end-of-program ❐ value added

#	Criterion	Yes	No	Comments
	Test Development & Professional Standards			
1	Has a professional organization developed the assessment?			
2	Does the professional organization have a set of professional competencies or credentialing standards for professional practice?			
3	Is the professional set of standards available for review?			
4	Is the assessment congruent with the professional set of standards?			
5	Does the test blueprint compare favorably with that of a NOCTI assessment in the same area ensuring that similar areas are being covered to the same depth?			
6	Are the professional organization's competency measures congruent with the test emphases ?			
7	Is the intended use of the assessment clear, i.e., as (1) an end-of-program assessment based on a set of skills and competencies that are broad based or as (2) a rigorous, value-added credential intended to discriminate the few top individuals?			
8	Is the test's reading level appropriate for the intended industry or job?			
9	Is the same test used in industry?			
10	Are results of national pilot tests or field tests available?			
11	Are the tests recent or is there a schedule of test revision?			
12	Are multiple tests needed to cover the entire content of the trade?			
	Technical Analysis			
13	Is there a technical manual for the test that PDE can review?			
14	Are there alternate forms of the test?			
15	Is there a performance component to the test?			
16	Is there a criterion-referenced benchmark?			
17	Are there results of a validity study performed on the test?			
18	Is there a reliability coefficient?			

#	Criterion	Yes	No	Comments
19	Is there the same number of choices for each item?			
20	Are the choices for each item plausible?			
21	Are the items independent of each other?			
22	Is there an item-to-total (point biserial) Correlation?			
23	Is this test comparable to other national tests?			
24	Is the test listed/reviewed in Buros Institute's *Tests in Print, Test Reviews Online,* or *Mental Measures Yearbook*?			
	Test Administration & Security			
25	Does the test administrator maintain appropriate test security?			
26	Is there a time limit for the test?			
27	Are there appropriate accommodations for individuals with disabilities?			
28	Can the written component be administered on-line?			
29	Is the performance component to be completed in a specific time limit?			
30	Is the performance evaluated by an independent content expert?			
31	Is there a scoring rubric for each performance task?			
32	Is there a criterion-referenced benchmark for each performance task?			
33	Is there an established competency (cut) score for the assessment?			
34	Is the test pass/fail?			
35	Are students given feedback on their partial subject / skill knowledge?			
	Reporting			
36	Will the testing organization provide an annual report to PDE by Sept 1 with a summary of Pennsylvania student performance disaggregated by specific subgroups?			
37	Does the test organization provide both a detailed individual student report and a summary report to the school?			
38	Are the school and individual student report formats available for review?			

Test Evaluator: ____

Based on the information available, please make a recommendation or give comments about the adequacy of this assessment's content to assess job readiness and the technical information to determine validity.

Comments (specify the item):

Recommendations:

Appendix C: Evaluation Matrix for Occupational Competency Test Approval

Test Provider: **Test Tile:** **Test Evaluator:**

	Criterion	Possible Score	Item Score	Criterion Score	Test Score
					200
	Test Development & Professional Standards			75	
1	Has a professional organization developed the assessment?	10			
2	Does the professional organization have a set of professional competencies or credentialing standards for professional practice?	10			
3	Is the professional set of standards available for review?	10			
4	Is the assessment congruent with the professional set of standards?	10			
5	Does the test blueprint compare favorably with that of a NOCTI assessment in the same area ensuring that similar areas are being covered to the same depth?	10			
6	Are the professional organization's competency measures congruent with the test emphases ?	5			
7	Is the intended use of the assessment clear, i.e., as (1) an end-of-program assessment based on a set of skills and competencies that are broad based or as (2) a rigorous, value-added credential intended to discriminate the few top individuals?	4			
8	Is the test's reading level appropriate for the intended industry or job?	3			
9	Is the same test used in industry?				
10	Are results of national pilot tests or field tests available?	10			
11	Are the tests recent or is there a schedule of test revision?	3			
12	Are multiple tests needed to cover the entire content of the trade?				
	Technical Analysis			50	
13	Is there a technical manual for the test that PDE can review?	10			
14	Are there alternate forms of the test?	2			
15	Is there a performance component to the test?	10			
16	Is there a criterion-referenced benchmark?	2			
17	Are there results of a validity study performed on the test?	10			
18	Is there a reliability coefficient?	3			
19	Is there the same number of choices for each item?	2			

20	Are the choices for each item plausible?	**3**			
21	Are the items independent of each other?	**2**			
22	Is there an item-to-total (point biserial) Correlation?	**2**			
23	Is this test comparable to other national tests?	**2**			
24	Is the test listed/reviewed in Buros Institute's *Tests in Print, Test Reviews Online*, or *Mental Measures Yearbook*?	**2**			
	Test Administration & Security			**30**	
25	Does the test administrator maintain appropriate test security?	**10**			
26	Is there a time limit for the test?	**3**			
27	Are there appropriate accommodations for individuals with disabilities?	**3**			
28	Can the written component be administered on-line?	**1**			
29	Is the performance component to be completed in a specific time limit?	**3**			
30	Is the performance evaluated by an independent content expert?	**10**			
	Scoring			**25**	
31	Is there a scoring rubric for each performance task?	**10**			
32	Is there a criterion-referenced benchmark for each performance task?	**10**			
33	Is there an established competency (cut) score for the assessment?	**5**			
34	Is the test Pass/Fail?				
35	Are students given feedback on their partial subject / skill knowledge?				
	Reporting			**20**	
36	Will the testing organization provide an annual report to PDE by Sept 1 with a disaggregated summary of Pennsylvania student performance?	**10**			
37	Does the test organization provide both a detailed individual student report and a summary report to the school?	**5**			
38	Are the school and individual student report formats available for review?	**5**			
	Minimum score of 132 is necessary for approval				

	MUSTS:		**ACCEPTABLE IF PROMISED**		**TOLERABLE IF ABSENT**	
Development	2, 3, 4, 10	**[40]**	1, 5, 6, 8, 11	**[31]**	7	**[4]**
Technical	13, 15, 17, 22	**[32]**	14, 16, 18, 23	**[9]**	19, 20, 21, 24	**[9]**
Administration	25, 30	**[20]**	26, 27, 29	**[9]**	28	**[1]**
Scoring	31, 32	**[20]**	33	**[5]**	none	
Reporting	36, 37, 38	**[20]**	none		none	