A Calculus of Occupational Skill Attainment:
Building More Validity into a Valid Assessment System

Dr. Paul Munyofu
Pennsylvania Department of Education
Dr. Richard Kohr
Independent Assessment Consultant

Abstract

This study investigated several aspects of occupational skill assessment as implemented in one state: (1) What is the extent to which student achievement on the cognitive component was related to their achievement on the psychomotor component of the technical skill assessments? (2) How efficiently was their overall composite attainment calculated? And (3) How well did this attainment predict student productivity on the job as determined by the employer's customer satisfaction? A sample of 118 student attainment scores on the written and performance components showed positive correlation. Further, this attainment was positively correlated with employers' customer satisfaction ratings. The panel of 16 national experts who participated in this study concluded that the Nedelsky (1974) method used to set the cut score needed to be re-evaluated. They also recommended that the scheme of calculation for determining one composite achievement level from the two test components should be modified.

Dr. Munyofu is a Research Associate in Pennsylvania Department of Education. He can be reached at pmunyofu@state.pa.us.
Dr. Kohr is a retired Educational Measurement and Evaluation Supervisor in Pennsylvania Department of Education. He can be reached at RKohr@itech.net.

Introduction

The advent of the Carl D. Perkins Vocational and Technical Education Act of 1998, and the No Child Left Behind Act of 2001, ushered in a new era of educational accountability for career and technical education. With the passage of the Workforce Investment Act of 1998, states receiving Perkins funds were required to report to the United States Department of Education and the Department of Labor the extent to which these states were helping their students attain skills necessary for entry level employment and postsecondary education. States were also required to establish a system to report levels of student achievement of technical skills. While many approaches were available for reporting skill attainment, the Pennsylvania Department of Education (PDE) chose to utilize tests from the National Occupational Competency Testing Institute (NOCTI). These were occupationally specific, aligned to Classification of Instructional Programs (CIP) codes, developed in conjunction with industry, and were designed to measure entry-level job-ready attainment.

Career and Technical Education in Pennsylvania

The history of career and technical education in the state of Pennsylvania and the nation is a long one. By the mid-1880s vocational education in the form of industrial education was synonymous with institutional programs for youths. The children of defeated Native American leaders were sent to the Carlisle Pennsylvania Indian School, and the curriculum was job training (Clarke: Federal Education Policy & Off-Reservation Schools 1870-1933; a presentation of the Clarke Historical Library. Online at http://clarke.cmich.edu/indian/treatyeducation.htm). Both Vo-

Tech high schools and community colleges all across Pennsylvania received much support from federal funds. (Pennsylvania State Archives, RG-22 Records of the DEPARTMENT OF EDUCATION AGENCY HISTORY, from http://www.phmc.state.pa.us/bah/DAM/rg/rg22ahr.htm).

The focus of a national legislative movement was to properly equip secondary and postsecondary youths with the necessary tools that facilitate meeting the demands of emerging industries. If the United States is to remain at the forefront of the high-tech global marketplace, the workforce must possess the requisite technological competencies and academic skills (Education Encyclopedia, 2007). The legislative acts, popularly known as Perkins of 1984, Perkins II of 1990, Perkins III of 1998 and Perkins IV of 2006 further emphasized the new focus of career and technical education. Students who complete an approved career and technical education program are expected to be ready for postsecondary education and work.

> "The purpose of this Act is to develop more fully the academic and career and technical skills of secondary education students and postsecondary education students who elect to enroll in career and technical education programs, by-
>
> > (1) building on the efforts of States and localities to develop challenging academic and technical standards and to assist students in meeting such standards, including preparation for high skill, high wage, or high demand occupations in current or emerging professions;" (Carl D. Perkins Career and Technical Improvement Act of 2006, Sec. 2. (Purpose (1).

Part of the Act required eligible agencies to submit a Consolidated Annual Report (CAR) that included "Student attainment of career and technical skill proficiencies, including student achievement on technical assessments, that are aligned with industry-recognized standards, if available and appropriate" (113(b)(2)(A)(ii)). The assessments of occupational skill attainments are expected to meet the Perkins "Gold Standard." This is a reference to:

> a classification of technical skill assessments that the U.S. Department of Education, Office of Vocational and Adult Education, views as the most valid and reliable measurement of technical skill attainment. Specifically, the Gold Standard encompasses (1) technical skill assessments, developed by external, third-party agencies to assess national or state-identified standards (e.g., nationally validated employer/industry and postsecondary cluster standards); (2) national, state, or industry-developed credentialing or licensing exams, typically used to control entry into a profession; or (3) standardized statewide assessments of technical skills created by state administrators for local agency use (DTI Associates, 2007, p. 5).

The National Occupational Competency Testing Institute

Even before the passage of the Carl D. Perkins Vocational Act in 1963, Pennsylvania supported a loosely organized system of student occupational competency testing (Walter, 1984). With the Act, more students were enrolled in vocational programs that demanded a more organized system of assessing competency (Walter and Kapes, 2003). It was generally agreed that printing, distributing, administering, and scoring of examinations imposed an impractical burden on limited state resources. A consortium of 23 states ardently expressed that a third-party, nationally coordinated effort was needed to develop occupational competency examinations~~,~~ in order for honest validation, establishing reliability, and other necessary construct measures. Most importantly, even the leading test development states were unable to experiment or carry on essential research, test development, field-testing, continuous revision, giving feedback to the states, and providing important test results and comparative, qualitative data. It was clear there was a need to professionally coordinate national efforts through a voluntary consortium effort (National Occupational Competency Testing Institute history online, from http://www.nocti.org/History.cfm). To that end, NOCTI became well established. Now NOCTI also owns a newly formed for-profit subsidiary, The Whitener Group, Inc., that provides a variety of assessment services for business and industry.

NOCTI has become a leading provider for occupational competency end-of-program assessments and services (NOCTI, 2007; Munyofu, 2007). By joining NOCTI, Pennsylvania gained the benefits of the national effort to produce quality occupational competency testing instruments to determine job-readiness among graduates of career and technical education programs. These tests were norm-referenced. Member states had the flexibility to choose how they interpreted the test results. Pennsylvania's initial choice was to identify students who performed at or above the national norm. These students were at that time considered as having distinguished themselves. They were awarded the governor's Pennsylvania Skill Certificate. Several unanswered questions remained. How did one know that an individual among the top half of those tested was good enough to be hired? (Munyofu, 2007, p. 4)

The Occupational Tests

The NOCTI tests are designed to measure both cognitive and psychomotor domains of career and technical education. The written component of approximately 200 multiple-choice items covers the entire program as outlined in the corresponding Classification of Instructional Programs (CIPs) of about 120 competencies. A written test may take approximately two to three hours. The performance component, on the other hand, consists of two to seven “jobs” which collectively address maybe 30 to 40 of the 120 competencies. This portion takes from three to four hours to complete.

The Performance (Psychomotor) Tests

Performance assessments consist of a series of tasks that make up a job. Individuals are required to complete jobs based on instructions provided in the test administration guidelines. Individual performance is rated by respective industry practitioner evaluators using specific criteria provided in the assessment’s evaluator guide. The evaluator selects the rating that best defines the work being completed. Some tasks have five options (A-E). Others have a combination of options (A-C-E or A-E). The evaluator is only allowed to rate the individual with the ratings that are provided. Evaluator directions include the criteria for determining the process used and the results (product) achieved, including the value for each criteria based on a particular point scale.

In Computer Networking Fundamentals (excerpted from one of NOCTI’s Technical Manual), for instance, the student might be required to:

> Create simple LAN with three PC’s, using an Ethernet hub or switch and three straight-thru cables to connect

workstations. Select the appropriate cable(s). Connect cable(s) to Network Interface Card (NIC) and hub or switch. Configure the networks settings. Check network connectivity and demonstrate file sharing.

Configuring the network might be rated by:

A = Participant properly configures the IP address;

B = Participant properly configures 2 of the 3 settings;

C = Participant properly configures 1 of the 3 settings;

D = Participant properly locates the network settings;

E = Participant did not configure or locate the settings, or did not complete.

If the task is utilizing a 25-point scale, then A = 25, B = 20, C = 15, D = 10 and E = 5. On checking network connectivity, which is in a 10-point scale, A = 10, C = 6, E = 2. Connecting cables to Network Interface Cards is rated on a 5-point scale with A = 5 and E = 1.

The Standards

Pennsylvania Department of Education (PDE) reports student performance on these occupational assessments as advanced, competent, basic and below-basic with the following descriptions:

Advanced Level – This level reflects mastery of competence and understanding of academic/career and technical skills and knowledge required for advanced placement in employment and/or postsecondary education.

Competent Level – This level reflects a solid acquisition of academic/career and technical skills and knowledge required to enter employment and/or postsecondary education.

> *Basic Level* – This 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 Level* – This level reflects a partial acquisition of skills and knowledge needed to perform a given assignment, task or operation on the 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."

Setting Cut Scores: The Nedelsky Method

The Nedelsky (1954) method of setting cut scores is used only with multiple-choice tests. It requires an expert judgment about the distracter of each test item. The judge's task is to look at the question and identify the answers that a minimally competent test taker would be able to recognize as obviously wrong, before resorting to guessing on the remaining choices. Livingston and Zieky (1982) used the following example from a test of language skill. The test taker's task is to choose the word or phrase that best completes the sentence.

> "My music teacher thinks that Marian Anderson
> sings_______any other contralto he has ever heard."
>
> (A) more well than (B) better than (C) the best of (D) more better over.
>
> A judge might decide that the borderline test taker would be able to eliminate wrong answers A and D.

> But the judge might decide that the choice between wrong answer C and the correct answer B is too difficult for the borderline test taker. The judge would then identify answers A and D as being so clearly wrong that the borderline test taker would be able to recognize them as wrong~~.~~ (p. 12).

When no choice is eliminated the candidate has a probability of guessing an answer correctly as 1 out of 4, hence *p*-value = 0.25. When 1 choice is eliminated, that probability is 1 in 3 or $p = 0.33$. Eliminating 2 choices leads to $p = 0.50$. When 3 choices are eliminated $p = 1.00$. The sum of the reciprocals over all the test items denoted the probable passing percent score for a single judge. The mean over all the judges is the percent cut score for the test.

For this method to provide valid and reliable results, the judges selected must be thoroughly knowledgeable about the subject matter for which the cut score is being developed. The panel must be sufficiently trained in this process so as to focus solely on the minimally competent candidate throughout the exercise. This training should include sufficient examples and discussion in order to increase inter-rater reliability.

Some researchers (Livingston and Zieky, 1982; Kapes and Welch, 1985) offered variations on the process, having compiled the judges' ratings. Some recommended using the median of the judges' ratings. Some suggested using a number halfway between the mean and median calculations. Others suggested eliminating the highest and the lowest score and calculating the mean of the remaining judges. Yet others allowed for adjusting the cut score using the mean minus a multiple of the estimated standard error of measurement ($S_E = s\sqrt{1 - r_{xx}}$) where s is the standard deviation of the scores and r_{xx} is the reliability index.

Should the judges make their judgments individually or try to reach a consensus? The method seems to work fairly well either way, if the number of judges is not too large. But even with a small number of judges, it may take some time to get a consensus on each test question, and with more judges, it will be even harder to get them to agree. Yet, the judges can make more valid judgments if they share information and opinions with each other.

One limitation of this procedure is that it requires all the judges to make their judgments at the same time and place. Another limitation is that, even with the shortcut, it is fairly slow (though not nearly as slow as trying to get a group consensus on each question). For either of these reasons, some choose to have the judges make their judgments individually, without communicating with each other. The state of Pennsylvania went so far as to allow the subject matter experts to make their judgments online, after a thorough face-to-face training, practice and discussion.

Livingston and Zieky (1982) also addressed additional considerations on the process by which judgments are made:

> One important issue in the application of Nedelsky's method (and of Angoff's and Ebel's methods) is whether or not to tell the judges the correct answers to the test questions. Giving the judges the correct answers may make the questions seem easier than they are and, therefore, bias the judges in the direction of a higher cut score. If you do not give the judges the correct answers, they may judge some of the correct answers to be wrong answers that a borderline test taker would eliminate, but this information can be valuable. If several judges eliminate the correct answer to the same question, that question may be defective. And if one judge eliminates many of the correct answers, that judge may be unqualified.

> However, if you do not give the judges the correct answers, the judges may feel that they are being tested and may forget that their judgments are supposed to indicate the responses of a borderline test taker. In addition, the judging process will surely take longer if the judges have to take the extra step of figuring out the right answer to each question. A good solution, if your situation permits it, is to have the judges take the test before the judging session and then give them the correct answers to use while they are actually making their judgments. (p. 13).

Other cut score setting methods had been considered when Pennsylvania initially chose to establish criterion-referenced benchmarks. Walter and Kapes (2003) compared alternate methods of setting Pennsylvania's cut scores on the NOCTI assessments. They described how Nedelsky compared against Angoff (1971), Ebel (1972) and Jaeger (1982).

The Problem

The state of Pennsylvania's Department of Education, Bureau of Career and Technical Education, has stressed the importance of a skilled workforce that will meet the demands of the future. Graduates are expected not only to know about welding but also to demonstrate that knowledge by actually welding. They are expected to be ready not only for work but also for postsecondary and advanced education and training. Pennsylvania demands that a graduate's Certificate of Competency or Pennsylvania Skill Certificate be a credential that attests to knowledge and skills the employer expects.

While the state has maintained such a high standard, several issues about their assessment system needed to be examined. Do students perform equally well on the written and the performance components of the test? If they do not, apart from accounting for individual differences and learning styles, how does one calculate a composite overall student attainment? The system of determining the overall level of attainment has been recently criticized as being too severe. Some critics claim that Pennsylvania should put more weight on the practical component of the end-of-program tests than on the written. That way when a student is advanced on the performance and competent on the written portion of the test, that student should be considered advanced on the whole test. A student who is advanced on one part and basic on the other should be, at the minimum, competent. The other half of the conversation, interestingly enough, would like extra weight added to the written component! When preparing a test specification blueprint for Heating, Ventilation and Air Conditioning (HVAC) one participant disagreed with this, commenting that:

> As an industry person in HVAC (Heating, Ventilation and Air Conditioning), I see the emphasis on written tests as counter to my world. As we spoke, after I show a new person how do a task, I ask them to show me they can do it, not give them a pop quiz. We need a hands-on assessment task list. I believe that performance is 60%, the written is 40%. I understand that some may see the performance portion as subjective, but let me assure you that in my world that is far from the truth (participant at a session to create a test specification blueprint, 2008).

Even more important is the issue of predictive validity for the assessment. Although the assessments are constructed in conjunction with industry, and industry representatives actually evaluate students' performance on the hands-on component, no empirical study has been conducted to see if there is a relationship between assessment scores and on-the-job performance. Customer satisfaction assessment needs to be a hallmark of an effective career and technical education program. This study was undertaken to address the following questions related to student technical skill attainment:

1. Is there a relationship between student achievement on the written and the performance components of the tests?
2. Is there a relationship between students' achievement on the tests and their future performance on the job as measured by their supervisors?
3. Is the scheme of calculation used to create a composite attainment level from the written and performance components efficient and sound?
4. Is the Nedelsky (1954) method of setting cut scores as currently applied in Pennsylvania appropriate, efficient and useful for determining competency in occupational skill attainment?

Methodology

In order to determine predictive validity for the assessment system, a questionnaire (see Appendix) was prepared and sent to all career and technical education school directors in the state. They were asked to solicit customer satisfaction information about some of their graduates from the employers who were in a position to evaluate their on-the-job performance. The school representatives would then return the questionnaire with the desired information about their graduates. For each graduate they would indicate the graduate's achievement on the written and performance components of the test, whether the graduate is employed in an area related to the field of study, and the level of employer satisfaction indicated on an accompanying Likert scale. The returned questionnaires by 17 schools contained data on a sample of 118 currently employed graduates from career and technical education.

Three years of trend data for 2005, 2006, and 2007 (Tables 2 – 4) was assembled and analyzed to determine if there was a correlation between student attainment on the written and performance components of the tests. The four tables and background information were sent to a panel of 18 nationally recognized measurement authorities with a request to assist in improving the system of determining over-all student occupational skill attainment on the basis of written and performance scores:

- Should the performance component carry the same weight as the written component?
- How do you interpret the data in tables 2, 3 and 4?
- Is it necessary to modify the attainment calculus?
- Would you suggest how such a modification might be accomplished?

The Cut Scores

To determine a student's achievement on the performance component, fixed cut scores of 80%, 75% and 70% were established at the onset of this reporting system. This determination was made through consultation with career and technical education instructors, industry representatives, a test provider of occupational skill assessments, and a measurement consultant contracted for the assessment project (Kapes, 2001; Walter and Kapes, 2003). Also at that time there was no obvious objective method for setting a cut score for this type of assessment. The written component was routinely criterion-reference benchmarked by a team of industry practitioners using the Nedelsky method (1954). With the competent level thusly initially determined, the basic level was calculated by subtracting five (5) percentage points from the competent level. The advanced level was calculated as five (5) percentage points above the competent level. No adjustments are made to these cut scores utilizing the Standard Error of Measurement (SEM) or the introduction of actual student performance on the tests (Munyofu, 2008; Kapes & Welch, 1985; Walter & Kapes, 2003).

An over-all occupational skill performance on these end-of-program assessments is determined for the purpose of reporting on Perkins accountability indicators. The final attainment level is the lower of the two scores. The bivariate function is:

$$(1) \qquad f(x,y) = \begin{cases} x, x \le y \\ y, y < x \end{cases}$$

That calculus for determining an overall composite attainment is depicted in the chart below (Table 1). A student who had Advanced (A) on the written, and Basic (B) on the performance was Basic (B) on the overall attainment. A student who had Below-Basic (BB) on the written and Competent (C) on the performance was Below-Basic (BB) on the overall attainment. Table 1 shows the bivariate functioning.

Table 1. Occupational Attainment Calculus

f	Achievement on Performance			
Written	A	C	B	BB
A	A	C	B	BB
C	C	C	B	BB
B	B	B	B	BB
BB	BB	BB	BB	BB

Historical Data

Over the previous three testing cycles (Tables 2, 3 and 4), student performance on the two portions of the NOCTI tests followed the accompanying pattern. The total number in the table consists of only those students who took the complete test, having finished the written and performance components of the tests. Students omitted from the data took only the written component, only the performance component, or parts of each. Of all 9743 students (Table 1) who were Advanced on the performance component: 4994 were also Advanced on the written, 1285 were Competent on the written, 1892 were Basic on the written, and 1572 were Below-Basic on the written.

Table 2. 2007 Bivariate distributions of scores on the two components

	Achievement on the Performance Portion				
Written Achievement	A	C	B	BB	Total
A	4494	234	158	1364	6250
C	1285	89	64	382	1820
B	1892	184	134	777	2987
BB	1572	183	138	917	2810
Totals	9743	690	494	3440	13867

Table 3. 2006 Bivariate distributions of scores on the two components

	Achievement on the Performance Portion				
Written Achievement	A	C	B	BB	Total

A	5039	298	206	1687	7230
C	1314	123	94	547	2078
B	1864	244	150	950	3208
BB	1266	169	127	1254	2816
Totals	9483	834	577	4438	15332

Table 4. 2005 Bivariate distributions of scores on the two components

	Achievement on the Performance Portion				
Written Achievement	A	C	B	BB	Total
A	6060	436	309	1910	8714
C	1093	127	89	570	1879
B	1322	212	166	1133	2833
BB	741	133	134	1359	2367
Totals	9216	908	698	4972	15793

Results

An SPSS Crosstabs analysis of the customer satisfaction data is given in Table 5. The related Chi-Square tests are given in Table 6. The results indicated that there is a significant correlation between achievement on the written tests and achievement on the performance components of the tests $\square^2(9, N = 118) = 76.246, p < .001$. Analyses were also conducted to determine the relationship between predictor variables (written and performance) and customer satisfaction. The analysis outputs are shown in Tables 7 – 10. Written correlation indices with Satisfaction (phi, Cramer's V, contingency coefficient) were statistically significant $\square^2(9, N = 118) = 20.696, p = .014$. However the Performance indices were not statistically significant $\square^2(9, N = 118) = 15.228, p =$

.085. The Written attainment is a better predictor of customer satisfaction after graduation than attainment on the Performance component.

Table 5. Attainment on the Written and Performance Tests

Crosstabulation							
			P				
			1.00	2.00	3.00	4.00	Total
W	1.00	Count	6	0	3	10	19
		% within W	31.6%	.0%	15.8%	52.6%	100%
		% within P	66.7%	.0%	21.4%	11.4%	16.1%
		% of Total	5.1%	.0%	2.5%	8.5%	16.1%
	2.00	Count	1	5	0	6	12
		% within W	8.3%	41.7%	.0%	50.0%	100%
		% within P	11.1%	71.4%	.0%	6.8%	10.2%
		% of Total	.8%	4.2%	.0%	5.1%	10.2%
	3.00	Count	0	0	10	15	25
		% within W	.0%	.0%	40.0%	60.0%	100.0%
		% within P	.0%	.0%	71.4%	17.0%	21.2%
		% of Total	.0%	.0%	8.5%	12.7%	21.2%
	4.00	Count	2	2	1	57	62
		% within W	3.2%	3.2%	1.6%	91.9%	100.0%
		% within P	22.2%	28.6%	7.1%	64.8%	52.5%
		% of Total	1.7%	1.7%	.8%	48.3%	52.5%
Total		Count	9	7	14	88	118
		% within W	7.6%	5.9%	11.9%	74.6%	100%
		% within P	100.0%	100.0%	100.0%	100.0%	100%
		% of Total	7.6%	5.9%	11.9%	74.6%	100%

Table 6. Chi-Square Indices on Written and Performance Attainment

	Value	df	Asymp. Sig.(2-sided)
Pearson Chi-Square	76.246[a]	9	.000
Likelihood Ratio	58.435	9	.000
Linear-by-Linear Association	19.865	1	.000
N of Valid Cases	118		

a. 11 cells (68.8%) have expected count less than 5. The minimum expected count is .71

Table 7. Written Attainment and Customer Satisfaction

			Crosstabulation				
			Satisfaction				
			1.00	2.00	3.00	4.00	Total
W	1.00	Count	0	3	9	7	19
		% within W	.0%	15.8%	47.4%	36.8%	100%
		% within Satisf	.0%	50.0%	27.3%	9.1%	16.1%
		% of Total	.0%	2.5%	7.6%	5.9%	16.1%
	2.00	Count	0	1	3	8	12
		% within W	.0%	8.3%	25.0%	66.7%	100%
		% within Satisf	.0%	16.7%	9.1%	10.4%	10.2%
		% of Total	.0%	.8%	2.5%	6.8%	10.2%
	3.00	Count	0	1	11	13	25
		% within W	.0%	4.0%	44.0%	52.0%	100.%
		% within Satisf	.0%	16.7%	33.3%	16.9%	21.2%
		% of Total	.0%	.8%	9.3%	11.0%	21.2%
	4.00	Count	2	1	10	49	62
		% within W	3.2%	1.6%	16.1%	79.0%	100.%
		% within Satisf	100.0%	16.7%	30.3%	63.6%	52.5%
		% of Total	1.7%	.8%	8.5%	41.5%	52.5%
Total		Count	2	6	33	77	118
		% within W	1.7%	5.5%	28.0%	65.3%	100%
		% within Satisf	100.0%	100.0%	100.0%	100.0%	100%
		% of Total	1.7%	5.1%	28.0%	65.3%	100%

Table 8. Chi-Square Indices on Written Attainment and Customer Satisfaction

	Value	df	Asymp. Sig.(2-sided)
Pearson Chi-Square	20.696[a]	9	.014
Likelihood Ratio	20.570	9	.015
Linear-by-Linear Association	7.310	1	.007
N of Valid Cases	118		

a. 9 cells (56.3%) have expected count less than 5. The minimum expected count is .20

Table 9. Performance Attainment and Customer Satisfaction

			Crosstabulation				
				Satisfaction			
			1.00	2.00	3.00	4.00	Total
P	1.00	Count	1	1	3	4	9
		% within P	11.1%	11.1%	33.3%	44.4%	100%
		% within Satisf	50.0%	16.7%	9.1%	5.2%	7.6%
		% of Total	.8%	.8%	2.5%	3.4%	17.6
	2.00	Count	0	1	1	5	7
		% within P	.0%	14.3%	14.3%	71.4%	100%
		% within Satisf	.0%	16.7%	3.0%	6.5%	5.9%
		% of Total	.0%	.8%	.8%	4.2%	5.9%
	3.00	Count	0	0	8	6	14
		% within P	.0%	.0%	57.1%	42.9%	100.0%
		% within Satisf	.0%	.0%	24.2%	7.8%	11.9%
		% of Total	.0%	.0%	6.8%	5.1%	11.9%
	4.00	Count	1	4	21	62	88
		% within P	1.1%	4.5%	23.9%	70.5%	100.0%
		% within Satisf	50.0%	66.7%	63.6%	80.5%	74.6%
		% of Total	.8%	3.4%	17.8%	52.5%	74.6%
Total		Count	2	6	33	77	118
		% within P	1.7%	5.1%	28.0%	65.3%	100%
		% within Satisf	100.0%	100.0%	100.0%	100.0%	100%
		% of Total	1.7%	5.1%	28.0%	65.3%	100%

Table 10. Chi-Square Indices on Performance Attainment and Customer Satisfaction

	Value	df	Asymp. Sig.(2-sided)
Pearson Chi-Square	15.228[a]	9	.085
Likelihood Ratio	12.468	9	.188
Linear-by-Linear Association	4.594	1	.032
N of Valid Cases	118		

a. 12 cells (75.0%) have expected count less than 5. The minimum expected count is .12

Should the performance component carry the same weight as the written component?

There was little consistency in the responses of the consultants. Three respondents (# 3, 7 and 16) thought that both components should carry the same weight. They recognized a business and industry's point of view that the up-coming workforce needs to realize that there are fixed standards that must be met for the individual to be economically viable in the workplace. Respondent #16 noted that the two components measure similar competencies. "One assesses students' abilities to answer questions about the competencies, an important skill since students must be able to communicate about their work. The other assesses students' abilities to implement the competencies, also very important."

Six respondents (# 2, 4, 5, 8, 12 and 13) indicated that they would like to see something other than equal weighting. One (#5) suggested that the performance should count more; another (#8) preferred the written. A third (#2) recommended that no decision should be made without data: "On the measurement side: A component that predicts the criterion best should have the most weight. Often one component predicts better than another. Further, components that have low reliability will predict less well than others and they should be weighed less. On the policy side: you would have to defend the choice based on solid evidence from job analyses rather than personal preferences of the authorities." In order to implement a compensatory approach, individual tests should be analyzed. Respondent #13 stated it this way. "Though many would argue that all jobs require significantly better cognitive skills than they did 20 years ago, all jobs are not the same. Establishing an equal rating for all occupations between cognitive and performance scores does not account for differences in these technical occupations. If you use an arbitrary weighting of the 2 measures without tying it to workplace reality it would be an unrealistic measure."

The rest of the responses were "maybe," or "unsure," or were neutral. Respondent #15 stated that "many methods of scoring can be used. But, there seems to be a need here to give weights to both the theoretical test as well as the practical test." Some of these are described in response to the last question below.

How do you interpret Tables 2 – 4?

If the correlations are high, respondents said, it means that the scores are highly related. If they are highly related then it suggests that there is a lot of redundancy in the testing, so that two separate tests may not be necessary. That is not the case according to the crosstabs analysis results (Tables 13 and 14).

According to Tables 11 and 12, the largest group scored A & A the next largest group scored A & B! If the written test was too easy or had test security been compromised, then one should pay more attention to the performance results as being more valid because they were generated through observing students actually finishing a task. A second observation was that the written achievement had continued to fall---the BB level was proportionately larger in each succeeding year. However, performance scores had risen. A third item was that the Competent Written score group was the smallest size of the written achievement groups on each table. Along with this was the very low number of students who score in the Competent and Basic levels on the performance tests. The data suggested that most students either can do very well or very poorly, with few students scoring in the middle two sections on the performance tests. The overall percent of candidates rated as Proficient OR Advanced, inclusively, is not unusual for certification exams of this nature.

A respondent observed: “We see somewhat of a trend from 2005 to 2007 in terms of increasing “A”s on the performance test (58% to 62% to 68%), whereas you don’t see that for the written (55% to 47% then steady at 47%). We also see a small trend indicating a decreasing number of people who get “A” on the written test but “BB” on the performance (12% to 11% to 9.5%), and an increasing number of people who get “A” on the performance test but “BB” on the written test (5% to 8% to 11%). Are teachers emphasizing hands on skills more but not the “academics” of the trade? Are evaluators trying to be more lenient in their scoring (e.g. not following the criteria as closely as they should)?

Table 11 Achievement Distribution over three years

Written and Performance Achievement percentage distribution of students		
2007	Written	Performance
A	0.47	0.68
C	0.13	0.05
B	0.21	0.03
BB	0.20	0.24
2006	Written	Performance
A	0.47	0.62
C	0.14	0.05
B	0.21	0.04
BB	0.18	0.29
2005	Written	Performance
A	0.55	0.58
C	0.12	0.06
B	0.18	0.04
BB	0.15	0.31

The statistical relationship between student performance level based on written and the practical performance evaluation was examined in analysis of the 2007 data. The results are presented in Tables 12, 13 and 14. Noteworthy is the rather low relationship between these two measures as indicated by the indices of association shown in Table 14.

Table 12. Attainment on the Written and Performance Tests for 2007

			Crosstabulation				
			Performance Test (PLP)				
Written Test (PLW)			1.0	2.0	3.0	4.0	Total
	1.0	Count	4494	234	158	1364	6250
		% within PLW	71.9%	3.7%	2.5%	21.8%	100%
		% within PLP	48.6%	33.9%	32.0%	39.7%	45.1%
		% of Total	32.4%	1.7%	1.1%	9.8%	45.1%
	2.0	Count	1285	89	64	382	1820
		% within PLW	70.6%	4.9%	3.5%	21.0%	100%
		% within PLP	13.9%	12.9%	13.0%	11.1%	13.1%
		% of Total	9.3%	.6%	.5%	2.8%	13.1%
	3.0	Count	1892	184	134	777	2987
		% within PLW	63.3%	6.2%	4.5%	26.0%	100.0%
		% within PLP	20.5%	26.7%	27.1%	22.6%	21.5%
		% of Total	13.6%	1.3%	1.0%	5.6%	21.5%
	4.0	Count	1572	183	138	917	2810
		% within PLW	55.9%	6.5%	4.9%	732.6	100.0%
		% within PLP	17.0%	26.5%	27.9%	26.7%	20.3%
		% of Total	11.3%	1.3%	1.0%	6.6%	20.3%
Total		Count	9243	690	494	3440	13867
		% within PLW	66.7%	5.0%	3.6%	24.8%	100%
		% within PLP	100.0%	100.0%	100.0%	100.0%	100%
		% of Total	66.7%	5.0%	3.6%	24.8%	100%

Table 13. Chi-Square Indices on Written and Performance Attainment

	Value	df	Asymp. Sig.(2-sided)
Pearson Chi-Square	268.760[a]	9	.000
Likelihood Ratio	266.199	9	.000
Linear-by-Linear Association	188.936	1	.000
N of Valid Cases	13867		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 64.84

Table 14. Written and Performance Correlation Indices

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Phi	.139			.000
	Cramer's V	.080			.000
	Contingency Coefficient	.138			.000
Interval by Interval	Pearson's R	.117	.009	13.839	.000[c]
Ordinal by Ordinal	Spearman Correlation	.122	.009	14.425	.000[c]
Measure of Agreement	Kappa	.065	.005	13.188	.000
N of Valid Cases		13867			

a. Not assuming the null Hypothesis
b. Using the asymptotic standard error assuming the null hypothesis
c. Based on Normal approximation

Is it necessary to modify the attainment calculus?

Based on the information provided, many of the participating experts were of the opinion that the calculus used to determine final skill attainment (Formula 1 and Table 1) was too stringent. "It seems to me," one expert (#2) stated, "that the procedure you are currently using for deciding who will pass is very arbitrary and should be studied in terms of how well people do on the job after taking the test or how well employers perceive these people are doing." In other words, doing a validity study using real job criteria. If you discover, for example that many people who do poorly on the real job receive "C" or better on your performance assessment, you would have evidence that your assessment is not valid." Expert #5 opined, "I do think the attainment calculus needs be modified. In particular I find the number of a/bb students unacceptable as such a discrepancy suggest to me the written assessment is measuring unrelated academic skills."

One respondent (#7) thought that there was no need to modify the attainment scheme. Another (#10), who chose not to commit one way or the other, commented that "The bottom line is that, you want the results to reflect your political objectives but I would not lower the percent from the written portion below what you already have." This was somewhat supported by #13, "The answer to this question really depends on the goal one is trying to achieve. However, we would recommend drilling down to at least the cluster level before making any kind of change in weighting. CTE's strength is in its connection to the workplace, so it is critical to maintain a metric that reflects that strength. One might compare what a change might do (if implemented) across the different clusters. Would it equate to more "A"'s in one group and less in another?"

How would you suggest such a modification be accomplished?

Many thought that the question was more political than not. They preferred to address cut score issues in the hope that the composite achievement problem will be indirectly resolved. One respondent offered the following refinement. "I would average the two levels and always round the results downward. So, some portions of the original Performance Calculus table would stay the same (e.g., A-A =A; A-C =C; C-B = B; B-BB = BB). And, others would change (e.g., A-B = C; C-BB = B; A-BB = B)." (See Tables 15 and 16.) The bivariate function would be

$$(2) \qquad f(x,y) = \left[\frac{x+y}{2} \right].$$

If Advanced = 4, Competent = 3, Basic = 2 and Below Basic = 1, then the function would be given by the chart below (Table 15).

Table 15. Modified Achievement Calculus

f	Achievement on Performance			
Written	4	3	2	1
4	4	3	3	2
3	3	3	2	2
2	3	2	2	1
1	2	2	1	1

Table 16. Modified Achievement Calculus

f	Achievement on Performance			
Written	A	C	B	BB
A	A	C	C	B
C	C	C	B	B
B	C	B	B	BB
BB	B	B	BB	BB

The calculation scheme proposed utilizes a form of compensation that would serve to safeguard against measurement errors, i.e. false negatives. The calculation would increase the proportion of students deemed at least Competent, a measure that would present the state's federal accountability results into a better light. Finally it would considerably reduce the proportion of students who are Below Basic.

Discussion

The results of this study indicate that there are serious issues that must be resolved before the student occupational skill assessment system in Pennsylvania can claim validity. This observation is in spite of the well-established credibility of the NOCTI Job-Ready assessments. It was commendable when Pennsylvania moved away from using the national norm as the standard for awarding the Pennsylvania Skills Certificate. They chose a criterion-referenced benchmarking model to determine whether a student who completed a career and technical education program was indeed ready for employment or postsecondary education.

When additional needs for information from the tests arose, the Pennsylvania assessment system did not evolve to accommodate these additional needs. These needs included: (1) benchmarks for the Advanced level in recognition of students who had distinguished themselves enough to be eligible for the Pennsylvania Skills Certificate; (2) criterion-referenced benchmarks for the Performance component of the tests; (3) benchmarks for the Basic level for those graduates who were employable, albeit needing additional training and remediation; and (4) evaluating the efficiency of determining overall student attainment.

The experts consulted in this study recognized that first and foremost, the benchmarking method needed to be updated. The Bookmark method (developed by CTB/McGraw-Hill, 1996) was suggested as the most appropriate for setting the three cut scores at the same time and applicable for both the written and performance components of the tests. "In general, the strengths of the Bookmark method are that it (a) accommodates constructed-response as well as selected-response test items; (b) efficiently accommodates multiple cut-scores and multiple test forms; and (c) reduces cognitive complexity for panelists" (Lin, 2006).

Other consultants suggested that Pennsylvania consider the Body Of Work model for setting the cut scores, as that method has been utilized for the Pennsylvania System of School Assessment (PSSA). However this would only be feasible for the written component. The performance (practical or hands-on) component focuses on the process as well as the completion of the assigned task. At this time neither Pennsylvania nor NOCTI has a system to preserve the body of work produced by the student. Yet it would be useful for test providers to consider investing in simulation programs to facilitate the assessments and preserve the testing process as well as the finished product.

NOCTI in 2008 started establishing national cut scores on their tests following the Pennsylvania model but with several modifications: (a) While in Pennsylvania the training of judges was conducted in a face-to-face format, the national training was conducted exclusively online. (b) Actual implementation of the judges' scoring was web based. (c) For each item the correct answer was already identified, so that the judges only needed to look at the item distracters and indicate which were obviously incorrect in the view of a minimally competent candidate. Of course this modification has the potential of tending towards higher cut scores (Livingston and Zeiky, 1982). (d) The highest and lowest judgments were dropped. Also dropped were judges who appeared not to follow the instructions correctly, in the opinion of NOCTI. (e) The Competent level was determined as the mean score for all the judges on the entire test, minus one standard error of measurement. The result was the percent of the items that must

be answered correctly for a student to attain the Competent level. Although NOCTI considered this adjustment as a means to establish more defensible cut scores, no empirical basis was offered. (f) The Basic level was 10 percentage points lower than the Competent level. The Advanced level was 10 percentage points above the Competent level. Again, the use of an arbitrary calculated range of ± 10 was not justified.
These modifications did not adequately address the concerns raised by the experts consulted in this study.

The first significant recommendation was that the state adopt a more up-to-date method for setting the cut scores. The second significant recommendation was that the calculus for determining overall attainment be modified in order to reduce the impact of possible false negatives. Often school administrators and career and technical education teachers advocate on behalf of some form of adjustment when a student achieved a much higher score on one form of the test than on the other. If the two scores cannot be reported separately then a variation of averaging the two scores appears to address that concern.

References

Angoff, W. H. (1971). *Norms*, scales, and equivalent scores. In R.L. Thorndike (Ed).Educational measurement (2nd ed.). Washington, D.C.: American Education on Education.

Carl D. Perkins Career and Technical Education Act. (2006). Carl D. Perkins Career and Technical Education Act of 2006 (Public Law 109-270). Washington, DC: U. S. Department of Education, Office of Vocational and Adult Education.

Clarke: Federal Education Policy & Off-Reservation Schools

1870-1933; a presentation of the Clarke Historical Library. Online at http://clarke.cmich.edu/indian/treatyeducation.htm

DTI Associates, Inc. (2007). Occupational Licensure: A Measure of Technical Skills Proficiency and More. A monogram submitted to Accountability and Performance Branch, Academic and Technical Education Division, Office of Vocational and Adult Education, U.S. Department of Education. Washington, DC.

Ebel, R. L. (1972). Essentials of educational measurement (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall. Education Encyclopedia online. (2007). Retrieved September 10, 2007 from http://www.answers.com/topic/history-of-vocational-and-technical-education

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. Educational Evaluation and Policy Analysis, 4, 41-475.

Kapes, J. T. (2001). Pennsylvania pass/fail cutoff scores for NOCTI written and performance exams based on 2001 national norm data. Report prepared for the Bureau of Career and Technical Education, Pennsylvania Department of Education.

Kapes, J. T., & Welch, F. G. (1985). Final report: Review of the scoring procedures for the occupational competency assessment program in Pennsylvania. University Park: Division of Occupational and Vocational Studies, The Pennsylvania State University.

Lin, J. (2006). The Bookmark Standard Setting Procedure: Strengths and Weaknesses. Alberta Journal of Educational Research vol. 52(1).

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.

Munyofu, P. M. (2007) Establishing a System to Evaluate Assessments of Student Occupational Skill Attainment. Online Journal of Workforce Education and Development, vol. II (4).

Munyofu, P. M. (2008) Differential Expectations of Student Performance on Occupational Skill Assessments Among Industry Practitioners: A Pennsylvania Example. Online Journal of Workforce Education and Development, vol. III (2).

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, *14*, 3-19.

National Occupational Competency Testing Institute. (2007). *A Brief History of NOCTI.* Online, from http://www.nocti.org/History.cfm

Pennsylvania Board of Education (1999) (22 Pennsylvania Code Section 49.17). Online. Retrieved from http://www.teaching.state.pa.us/teaching/lib/teaching/act48.pdf

Pennsylvania Department of Education. DEPARTMENT OF EDUCATION AGENCY HISTORY. Pennsylvania State Archives, RG-22 Records, from http://www.phmc.state.pa.us/bah/DAM/rg/rg22ahr.htm

Walter, R. A. (1984). An analysis of selected occupational competency assessment candidate characteristics and successful teaching. Unpublished doctoral dissertation, The Pennsylvania State University, University Park.

Walter, R.A. and Kapes, J.T. (2003) Development of a Procedure for Establishing Occupational Examination Cut Scores: A NOCTI Example. Journal of Information Technology Education, vol. 40 (3).

Zieky, M. and Perie, M. **A** Primer on Setting Cut Scores on .Tests of Educational Achievement. Online. Retrieved from http://www.ets.org/Media/Research/pdf/Cut_Scores_Primer.pdf

APPENDIX

Customer Satisfaction Survey

BCTE is interested in the extent to which student performance on occupational end-of-program tests is related to on-the-job performance. This is a part of an investigation about how accurately test cut scores help to predict success after graduation. The bureau will be able to modify how the cut scores are determined and consequently how student achievement will be used to evaluate career and technical education programs.

Please identify at most 8 of your former graduates who are employed and whose supervisor can provide you an evaluation of their job satisfaction. Then fill the table below with the student achievement on the written and performance portions of the NOCTI test. Please return this to me before September 30, 2008.

School:								
Student	**Employment**	**Test Results**		**Employer Satisfaction**				
Number	**Employed/ Related**	**Written**	**Performance**	**5**	**4**	**3**	**2**	**1**
0 example	Yes	C	A		√			
1								
2								
3								
4								
5								
6								
7								
8								

List students as 1, 2, 3, etc and no student names.

Is the student employed in a field related to the program completed? Indicate yes or no in this column.

What was the student's occupational achievement on the end-of-program tests, both written and performance?
A=Advanced, C=Competent, B=Basic, BB=Below Basic.

From the student's employer supervisor, please indicate the level of technical expertise demonstrated by the student on the job. Use 5=Very satisfied; 4=Somewhat satisfied; 3=Neutral; 2=Somewhat dissatisfied; 1=Not satisfied.

NOTES

The authors acknowledge the invaluable expertise and assistance from the following individuals.
Dr. Mariam Azin Pres Associates
Ms. Elizabeth Brown, National Association of Career and Technical Education Information

Dr. John Foster, National Occupational Competency Testing Institute
Dr. Kenneth Gray, Pennsylvania State University
Ms. Cynthia Gross, Pennsylvania Department of Education
Dr. Ronald Hambleton, University of Massachusetts
Dr. Aldo Jackson, Erie Career and Technical School
Dr. Leonard Lock, State University of New York at Plattsburg
Dr. Robert Mahlman, Ohio State University
Dr. Jim Masters, Independent Consultant
Dr. Anthony Nitko, University of Pittsburgh
Dr. John Townsend, National Association of Career and Technical Education Information
Dr. Richard Walter, Pennsylvania State University
Dr. Chester Wichowski, Temple University
Dr. Van Yidana, University of Rhode Island