DOCUMENT RESUME

ED 400 465 CE 072 863

AUTHOR	Stecher, Brian; And Others
TITLE	Using Alternative Assessments in Vocational Education.
INSTITUTION	Rand Corp., Santa Monica, Calif.
SPONS AGENCY	National Center for Research in Vocational Education, Berkeley, CA.
REPORT NO	DRU-1480-NCRVE/UCB
PUB DATE	Sep 96
NOTE	151p.; Prepared as a supplement to "Getting To Work: A Guide for Better Schools," a National Center for Research in Vocational Education (NCRVE) training package.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC07 Plus Postage.
DESCRIPTORS	Educational Change; Educational Development; Elementary Secondary Education; *Evaluation Methods; Instructional Improvement; Minimum Competency Testing; Performance Tests; *Portfolio Assessment; Portfolios (Background Materials); *Student Certification; *Student Evaluation; Student Organizations; Teacher Certification; *Vocational Education
IDENTIFIERS	*Alternative Assessment

ABSTRACT

This report describes alternative assessments in vocational education, reviews examples from extended case studies, and discusses criteria to use to choose among assessment alternatives. Chapter 1 is an introduction that contains brief summaries of each of the six assessment alternatives to familiarize the reader with the range of the sample and the variety of approaches that were represented. Chapter 2 examines the primary purposes served by assessments in education and the specific conditions that are creating pressure for alternative methods of assessment among vocational educators: the changing student population and the rapidly evolving skill mix that must be reflected in vocational programs. Chapter 3 describes the range of assessment methods, from common multiple-choice tests to new constructed-response alternatives, including performance tasks, senior projects, and portfolios. Chapter 4 discusses the quality and feasibility of alternative assessments. Chapter 5 identifies other factors relevant to choosing appropriate assessment strategies and the advantages associated with particular choices. Chapter 6 presents examples of the kinds of assessment decisions confronting vocational educators and shows how the results of this study can contribute to these decisions. Six appendixes describe each of the following case studies in detail: Kentucky Instructional Results Information System; Laborers-American General Contractors Environmental Training Assessment; Oklahoma's State Competency-Based Testing System; National Board for Professional Teaching Standards; Vocational Industrial Clubs of America National Conference--Job Skills Contests and Leadership; and Career-Technical Assessment Program. Contains 30 references. (YLB)

RAND

Using Alternative Assessments in Vocational Education

Brian Stecher, Mikala Rahn, Allen Ruby, and Martha Alt, with the assistance of Brian Ward

DRU-1480-NCRVE/UCB

September 1996

Prepared for the National Center for Research in Vocational Education/University of California, Berkeley

Preface

Student assessment has always played an important role in vocational education, and recent changes in assessment practices may hold great promise for vocational educators. The present study examines alternative forms of assessment in light of the needs of vocational educators. This report is the first of two products derived from the study. In the report, the authors describe alternative assessments, review examples from extended case studies, and discuss criteria to use to choose among assessment alternatives. These results should be of interest to educators at the state and local level, particularly those responsible for decisions about the form and use of assessment systems. The second product will be a set of training materials to help prepare vocational educators to make effective decisions about assessments. These materials will be prepared as a supplement to Getting To Work: A Guide for Better Schools, a recent NCRVE training package.

Contents

Tables

Figures

Acknowledgments

The initial framework for this study was extremely broad, as was our list of operational assessment programs to investigate. Fortunately, the members of our Advisory Committee helped us narrow our focus and select a interesting set of assessment case studies. For this help we are grateful to Charles Hopkins, Oklahoma Department of Education; Harry O'Neil, University of Southern California; Dean Peterson, Glendale Union High School District; Stanley Rabinowitz, EDWEST; Catherine B. Smith, Michigan Department of Education; Jonathan Tropper, Center for Research on Evaluation, Standards, and Student Testing, UCLA. We wish to thank the program administrators, teachers, participants, and researchers who responded to our interview questions as part of the case studies of the Kentucky Instructional Results Information System, the Vocational/Industrial Clubs of America, the Laborers-Associated General Contractors environmental training and certification programs, the Oklahoma Department of Vocational-Technical Education competency-based testing program, the National Board for Professional Teaching Standards, and the Career-Technical Assessment Program. In addition, there were many other staff from testing and assessment projects who helped us gather information in our initial round of investigations of operational alternative assessments. Finally, credit is due to David Adamson for his efforts to improve the clarity of the report and to Judy Wood for her help with document preparation.

1. Introduction

During the past few years, economic concerns have prompted a number of proposals to reform federal vocational education and employment training programs. Assessment plays a prominent role in these reform proposals. Reports about the inadequate skills of high school graduates, the rapidly changing demands of the work place, and the declining competitiveness of U.S. firms in the international marketplace, have stimulated a variety of proposals to change the organization and structure of employment preparation programs. The 104th Congress has argued at length about this issue, but as the election of 1996 approaches Democrats and Republicans have not been able to reach a consensus about the shape of future federal vocational education programs.

However, despite strong differences in their approach to reform, all sides seem to agree on the need for trustworthy methods for assessing students' skills. A noteworthy example of this convergence on assessment is the continuing debate about the relative importance of job-specific skills and more general industry-wide skills. Both sides in this debate—those who recommend a greater focus on broad industry skill standards at the secondary level (Boesel and McFarland, 1994) and those who place greater emphasis on occupational specific skills (Bishop, 1995)—agree about the need for a system to assess skills in reliable and valid ways. Almost all policymakers think it is essential to measure the degree to which students have mastered the skills around which training is focused.

Moreover, many vocational educators are advocating the wider use of alternative assessments, such as portfolios, exhibitions and performance events for measuring skills of either type. This interest in new measures derives in part from the changes that are occurring in vocational education. Educators and employers believe that the work world is changing and vocational education must adapt if it is to serve students well. The changes in the workplace are complex and not completely understood, but most believe that future employees will need integrated academic and vocational knowledge, broad understanding of occupational areas, the ability to interact creatively with their peers, and higher-order cognitive skills that allow them to be flexible, learn rapidly, and adapt to ever-changing circumstances. As a result, vocational training needs to place greater focus on integrated learning, critical thinking skills, and connections between vocational and academic skills, rather than rote mastery

of narrow occupation-specific skills that characterized vocational education in the past. This vision demands a major rethinking of the goals, organization, content and delivery of vocational education, as well as the manner in which students and programs are assessed.

The educational measurement community is engaged in an equally serious rethinking of the structure of assessment (Wolf, 1992; Mehrens, 1992; Wiggins, 1989). Traditional multiple choice methods are being criticized for a variety of reasons: they can lead to narrowing of curriculum; test preparation practices may inflate scores in high stakes situations; there are consistent differences in average performance between racial/ethnic and gender groups, etc. (Koretz et al., 1991; Shepard & Dougherty, 1991; Koretz et al., 1993; Shepard, 1991; Smith & Rothenberg, 1991). Many educators advocate the use of alternative approaches, including open-response items, realistic simulations, extended performance events, exhibitions, judged competitions, portfolios, and other forms of elaborated student demonstration.

Educators are working to find ways to improve the technical quality and feasibility of such performance-based assessments. On the positive side, the distinguishing feature of most alternative assessments is "authenticity," i.e., students perform an activity or task as it would be done in practice rather than selecting from a fixed set of alternatives. On their face these activities have greater validity than multiple choice tests because success is clearly related to the criterion of interest, be it writing, problem solving, or performing job tasks. On the negative side, students' performance is not as consistent from one task to the next as it is with multiple choice items, and the scores produced by alternative assessments are not as dependable or interpretable as those produced by traditional tests (Shavelson, et al., 1993; Koretz, et al., 1994). These issues are unresolved at present, but there appear to be trade-offs between cost, intrusiveness, dependability and interpretability.

The uncertainties surrounding vocational education and educational assessment provide the context for our inquiry into vocational education assessment. Research suggests that assessment can play an important role in systemic educational change such as that being envisioned for vocational education. The question we are exploring is: What forms of assessment might best meet the needs of vocational education and how can educators make intelligent choices among assessment alternatives? Our evaluation of assessment alternatives will pay particular attention to the purposes for which the assessment is used, the quality of the information provided, and the practicality or feasibility of the assessment approach.

Purpose and Procedures

This project has a two-fold purpose: (1) to provide information about promising educational assessment alternatives that may meet the needs of vocational educators, and (2) to develop materials to help vocational educators make better decisions regarding the use of alternative assessments. The present study addresses the first goal, i.e., to evaluate alternative assessments in the context of the current needs of vocational educators. To that end we gathered information about selected assessment systems, summarized it in a set of detailed case descriptions, reviewed it critically from the perspective of vocational education, and identified some of the important factors that will affect the choice of assessments in the vocational context. The next phase, which will be completed in a few months, will be to convert our knowledge into training and evaluation materials useful to vocational educators.

We began our investigation broadly, reviewing the literature and contacting experts in the field to look for promising examples of operational assessment systems that were applicable to vocational education. We developed a set of frameworks for organizing our thinking about the needs of vocational educators and for classifying types of assessment, uses of assessment, and dimensions of assessment quality. These ideas are reflected in discussion about the needs of vocational educators and the range of assessment alternatives which follow the description of the cases we studied. We then identified a tentative list of exemplary assessment alternatives—both within vocational education and in related education and training sectors. For each project we collected initial descriptive data from the printed record and from telephone interviews and compiled these into a working casebook.

A panel of expert advisors familiar with vocational education and assessment was formed to guide our work from both technical and practitioner perspectives and to select the assessment reform efforts that would be reviewed in depth. The panel met in March of 1995 and offered their advice about our selection of cases and plans for organizing information. To achieve our goal of providing vocational educators with relevant information and helpful procedures for selecting (or developing) alternative assessments, the panel members felt it was important to include a diverse set of assessments in our sample. They encouraged us to select assessment cases that differed in terms of their purposes and the uses of the results, the types of knowledge and skills being assessed, the types of assessment strategies being used, and the organization and structure of the assessment system.

With these factors in mind, we selected six cases for in-depth investigation:

- Career-Technical Assessment Program (C-TAP)
- Kentucky Instructional Results Information System (KIRIS)
- Laborers-AGC (Associated General Contractors) environmental training and certification programs
- National Board of Professional Teaching Standards certification program (NBPTS)
- Oklahoma Department of Vocational-Technical Education competency-based testing
- Vocational/Industrial Clubs of America (VICA) national competition

We developed a common set of questions to guide our examination of the six cases. The questions focused on description, implementation, administration, consequences, feasibility, quality, and applicability to vocational education. In each instance we gathered information to address these questions from a variety of sources, including descriptive materials provided by the program, the research literature, telephone interviews, and one- to two-day site visits. During the site visits (which included four of the six programs) we interviewed staff and observed assessment activities.

After the data were collected, we constructed a thorough description of each assessment, including the features we deemed to be most relevant to vocational education. One member of the research team assumed primary responsibility for each of the assessment activities. This person coordinated the data collection and was responsible for writing up the case summaries according to a common format. One person assumed an editorial role and rewrote the case summaries to provide greater consistency of presentation and voice. We were not formally evaluating each of the efforts, and did not attempt to reach a conclusive judgment about each program. When questions could not be answered from available sources we left them unresolved, and when there was disagreement among sources or we found contradictory information we reported the disagreements. Finally, we conducted an impressionistic review of the case reports looking for insights that would be relevant to vocational educators.

In the following sections we briefly describe the cases we studied. In subsequent chapters we use these cases to illustrate the range of concerns and choices that confront vocational educators. Specifically, we discuss the educational uses of assessment and the specific needs of vocational educators. Then we describe various types of measures that vocational educators might use and the factors

that might affect the choice of assessment alternatives, including the quality of the information provided, the feasibility of various options and a number of other issues that arose from the case studies. In each instance, we present information from the cases to illustrate the issues being discussed.

Our sample was both too small and too diverse to permit strong generalizations about the type of assessment to use in a particular situation. Instead it provided illustrations of a variety of trade-offs that confront the developers of educational assessments, trade-offs that are relevant to vocational educators, as well. We discuss these trade-offs, presenting illustrations drawn from the cases and relating these cases to the vocational education context. In the concluding section, we consider two prominent assessment challenges facing vocational educators, improving programs and certifying occupational mastery, and draw some implications from our study for selecting or developing assessments to support those functions.

A note on terminology: The collection of assessment activities we reviewed was quite diverse, creating minor problems in terminology. The sample ranges from developmental efforts (C-TAP) to fully operational testing programs (Oklahoma); from job specific measures (VICA) to broader occupational assessments (NBPTS), and from single tests (Oklahoma) to assessment systems (KIRIS). Because of this diversity, it is difficult to find simple terminology to refer to all these assessment efforts. They are not all "tests" in the traditional use of the word, nor are they all "testing programs." We will use the terms *assessment, assessment activity,* and *accountability system* to refer to our cases, in general. When discussing a specific case we may use a narrower, more focused term, such as *test* or *measure,* as appropriate.

Brief Description of Cases

The following paragraphs contain brief summaries of each of the six assessment alternatives to familiarize the reader with the range of our sample and the variety of approaches that were represented. More thorough descriptions of each of the six assessment activities are contained in appendices.

Kentucky Instructional Results Information System

The Kentucky Instructional Results Information System (KIRIS) is a statewide assessment system for elementary and secondary schools that is part of a major reform of Kentucky public education. The assessment system is designed both as an accountability tool and as a lever to promote changes in curriculum and

instruction. KIRIS uses multiple measures of achievement, including open-ended written questions, group performance events, and portfolios of students' best work to produce school-level scores. Also factored into a school's accountability score are noncognitive measures of attendance and retention. Significant rewards are attached to success and failure (schools can earn thousands of dollars for high performance and they face the threat of external intervention for continued failure to improve performance). Partially as a result, KIRIS has affected teachers' behaviors and brought about changes in schools that are consistent with the larger state reform effort. On the other hand, stakeholders have raised questions about the quality of the scores and the fairness of the awards. Independent evaluations have identified technical shortcomings that threaten the validity of the awards and of school comparisons. Kentucky educators are working to respond to these concerns and to improve the system.

Career–Technical Assessment Program

The Career-Technical Assessment Program (C-TAP) was originally developed by Far West Laboratory as part of a proposed state-wide certification system for vocational students in California. However, California has not implemented such a certification system, and C-TAP has evolved into a classroom assessment tool that is embedded in the curriculum with the purpose of improving instruction. It is currently being field tested for use in five career areas: agriculture, business, health careers, home economics, and industry and technology education. C-TAP contains three components: students are to complete a portfolio, a project and a scenario. The portfolio contains work samples and summaries of their best work, a writing sample, and evidence of generic work skills they have mastered (such as a resume, job application, and reference letter). The project is a long-term activity that varies by course; it is judged in terms of level of preparation, progress, final product and presentation of work. The scenario is a timed essay, written in response to a description of a realistic occupational situation in the vocational area being studied. C-TAP has been adopted on a teacher by teacher basis. As a result, adoption is not standardized, and there is variation in which components teachers use, how they interpret each component, and what use they make of the C-TAP results (e.g., they usually contribute to students' grades and may also be required to graduate from a class). The portfolio is the most widely adopted and appreciated part of C-TAP; teachers believe the work samples clarify whether the student understands the material. Teachers adopting C-TAP believe that it has improved instruction. The greatest benefit derives from students

preparing descriptions for their portfolios of the skills they have learned in the work samples. It is hoped that the portfolios also will provide students with a record of their abilities that can be used to impress potential employers and schools of higher education. Far West Laboratory continues to work on technical quality issues related to C-TAP with a current focus on standardizing scoring. In the future C-TAP may return to its original intended use, as part of a certification program; the state has already incorporated C-TAP into certain reform initiatives.

National Board for Professional Teaching Standards

The National Board for Professional Teaching Standards (NBPTS) offers voluntary national certification to recognize highly accomplished K–12 teachers, using a range of alternative assessment methods. The Board aims "to establish high and rigorous standards for what teachers should know and be able to do, to certify teachers who meet those standards, and to advance other education reforms," all with the underlying goal of improving student learning. The standards and tasks by which candidates are judged were developed mainly by other teachers. To obtain the NBPTS certificate, teachers prepare an extensive portfolio demonstrating their preparation, classroom work, teaching strategies, and professional activities, as well as participate in two days of performance activities at a regional assessment center. Standards committees (mostly teachers) use a multistage process to develop subject-matter standards, and Assessment Development Laboratories create the assessments. Assessments are still being developed/tested for many of the categories (combining one of 4 grade levels with one of 14 subjects), but in 1995-96, National Board certification was available in two categories. Extensive reviews of validity, reliability, and other quality-related factors have, on the whole, produced positive results. Areas that need improvement include reduction of the costs of test development, administration, and scoring, and better expression of directions to candidates for tasks (including expected length). Candidates who complete the process find it extremely rewarding despite the substantial burdens. The Board intends for the system to drive preservice and inservice training and even to influence state licensing standards.

Vocational/Industrial Clubs of America

Vocational/Industrial Clubs of America (VICA) is a national organization for secondary and postsecondary students in some 60 vocational/technical fields. VICA conducts national "contests" that focus on performing occupationally

specific skills in realistic contexts. Many of the skill areas include a written exam, as well. The national competitions are the culmination of local, regional, and state contests; winners proceed to the next level. Performance in the contests is judged by experienced industry practitioners using specific task-related criteria. The organization places high priority on fairness and consistency in judging; however, no research has been done on the validity, reliability, or equity of the test content or scoring methods. VICA aims for its contests to be closely tied to instruction in the relevant field, though this varies across competition fields and across instructors. Industry practitioners develop the performance tasks and the written tests, under VICA's guidance. This extensive industry involvement increases the relevance of the assessments to the workplace. Students and teachers gain a reality-based and up-to-date picture of the performance and skills expected in their industry from the involvement of practitioners. The written tests are primarily multiple-choice, but there are a few open-ended items as well. The VICA model would be relatively easy to replicate in schools. The most substantial obstacle would be recruiting experienced and knowledgeable industry people to design and judge the competitions.

Laborers–AGC Environmental Training and Certification Programs

The Laborers International Union of North America and the Associated General Contractors of America (AGC) cooperatively fund and manage a program of courses and assessments to train and certify environmental clean-up workers (and construction laborers). The courses, which are taught at affiliated local training schools, must comply with federal government regulations, which focus on avoiding potential threats to health and safety. The assessment system includes both performance events with real equipment (which take place multiple times during the course) and criterion-referenced multiple-choice tests (which occur at the end). The assessments are used to certify each individual's competence, as well as to monitor program success and report program completion information. In the last few years, Laborers-AGC has started to evaluate and strengthen the technical quality of its environmental assessments, though employers' evaluation of certified employees is already quite positive. Because the Fund is a shared venture between labor and management, employers have immediate input if their needs are not being met. The Laborers-AGC model carries high operational costs because of the depth and breadth of Laborers-AGC's hands-on activities and

the need for extensive space, (e.g., to create mock hazard sites), expensive equipment, and supplies actually used on the job.

Oklahoma Competency-Based Testing

The Oklahoma competency-based testing system encompasses a range of multiple-choice and performance-based assessments for both secondary and postsecondary students. The Oklahoma Department of Vocational-Technical Education (Oklahoma Vo-Tech) developed and oversees these tests, which are used to certify students for employment, to improve instruction and student learning through competency-based curriculum and assessment, and to report program improvement and accountability data at the state level. Students are required to pass two local performance assessments, attain all locally identified competencies, and then pass a written multiple-choice test. The responsibility for establishing competencies, certifying mastery, and conducting performance assessments rests with individual programs, with the associated variation from site to site. The multiple-choice component of the program is administered centrally and standardized across sites. Criterion-referenced multiple-choice tests have been developed for 190 occupational titles, which are categorized into 26 program areas. The tests measure occupation-specific knowledge and skills. State staff feel confident about the tests' content validity based on the strong employer input into the assessment system, but no formal validation research has been done. The curriculum guides are used inconsistently across schools and occupational areas, so for some classes instruction and testing are not very closely tied. Oklahoma has a long-standing tradition of centralized state control, university support, and substantial state funding for vocational education. With less funding and less acceptance of centralized authority, other states may be hard-pressed to follow Oklahoma's example.

Organization

The rest of this report is organized as follows: Chapter 2 examines the primary purposes served by assessments in education and the specific conditions that are creating pressure for alternative methods of assessment among vocational educators: the changing student population and the rapidly evolving skill mix that must be reflected in vocational programs. Chapter 3 describes the range of assessment methods, from common multiple-choice tests to new constructed-response alternatives, including performance tasks, senior projects and portfolios. Chapter 4 discusses the quality and feasibility of alternative

assessments. Chapter 5 identifies other factors relevant to choosing appropriate assessment strategies and the advantages associated with particular choices. Chapter 6 presents a examples of the kinds of assessment decisions confronting vocational educators and shows how the results of this study can contribute to these decisions. After the body of the report, six appendices describe each of the case studies in detail.

2. The Assessment Challenge Facing Vocational Education

The process of selecting or developing assessments begins with an examination of the potential uses of the information that is produced. These purposes drive the choice of assessment method. There are three broad uses for educational assessment: improvement of learning and instruction, certification of individual mastery, and evaluation of program success, and all three are relevant to vocational education. The first part of this chapter explores these purposes with illustrations from our case studies of alternative assessment systems.

Vocational educators also need assessments that are responsive to the specific demands of their field. For example, all the proposed federal reforms of vocational education require accountability of one form or another, and they assume vocational educators will produce measures of student or program performance that can fulfill this function. In addition, vocational assessments must be sensitive to the characteristic of the students being educated and the nature of the content they are supposed to understand. The second section explores recent changes in the context of vocational education and the implications of these changes for choosing assessments.

Purposes

Educational assessments can serve a variety of purposes, and the choice of assessment will depend in part on the manner in which the assessment information will be used. The Office of Technology Assessment identified three primary uses of assessment: 1) measuring student learning, 2) certifying mastery, and 3) providing program performance information (U.S. Congress, Office of Technology Assessment, 1992). All three purposes are relevant to vocational education. Vocational teachers use the results of tests and other assessments to monitor the progress of students, diagnose their needs, and make instructional plans. When students complete courses or sequences of courses, vocational programs use assessments to certify that students have achieved a required level of mastery or have met industry standards. Finally, aggregated information about student progress (acquired knowledge and skills, success in courses, etc.) is used to judge the quality of vocational programs. Although a

single assessment can be used for many purposes—for example, standardized test results are used by teachers to identify individual student weaknesses and target instruction, and they are used by legislators and the general public to judge the quality of the state education system—the same test may not be equally effective for all these purposes. Therefore, the choice of assessment should be made with these three potential uses of the information clearly in mind.

Teachers are usually responsible for measuring individual learning within the classroom and using this information to improve instruction and promote learning. Through direct observation as well as a variety of formal and informal assessment strategies, teachers keep track of what students learn, which instructional approaches work, and where changes need to be made. To be most helpful for improvement, assessments should provide detailed information on the specific knowledge and skills that have been taught in the class. They should be administered often and graded quickly, and information should be provided to teachers and students so that adjustments can be made. For the purposes of instructional improvement, assessments can be either on-demand or cumulative. Teachers administer, score and use them in conjunction with other knowledge of student performance, so less of a premium needs to be placed on technical quality.

Assessments can also be used to verify that students have mastered a particular set of skills or body of knowledge. Such information is used for the purposes of selection, placement, promotion and certification. Assessment for mastery may focus on general abilities (such as tests for college admission) or specific skills (such as for professional licensing). These decisions demand attention to the quality of the measures, including their reliability, validity, and fairness. Because these decisions have direct bearing on students' futures, they are often based on multiple rather than single measures.

Assessment can also be used to provide information about the quality of programs, schools, and districts that are providing education and training. This accountability may be based on individual performance or on group (e.g., class or school) performance. Because they are used to compare and reward programs, accountability assessments should demonstrate a high degree of reliability, and validity.

The case studies demonstrated the full range of purposes, as is illustrated in Table 1.

Table 1

Purposes of Assessment in the Sample

	Measuring Individual Learning for Instructional Improvement	Certifying Mastery	Holding Programs Accountable
California Testing Assessment Program (C-TAP)	✔		
Laborers-AGC environmental training		✔	
Kentucky Instructional Results Information System (KIRIS)	✔		✔
National Board of Professional Teaching Standards (NBPTS)		✔	
Oklahoma Department of Vocational and Technical Education	✔		✔
Vocational Industrial Clubs of America (VICA)	✔	✔	

Vocational Program Context

Vocational educators need assessments that are sensitive to the unique features of the vocational context. In particular, vocational educators face changes in the nature of the students enrolling in vocational courses and changes in the nature of skills being taught in those courses. Both features need to be understood to make wise assessment choices.

Vocational Student Population

Vocational education is offered in comprehensive high schools, area vocational-technical schools, community colleges, private proprietary schools and public technical institutions. Although recently there has been a movement to coordinate secondary and postsecondary vocational coursework through articulation agreements, for the most part, the secondary and postsecondary vocational education delivery systems remain separate. (The two are often most coordinated in area vocational-technical schools/centers where high school students and adults frequently enroll in the same courses.)

At the secondary level, there has been a decrease in vocational course-taking in favor of more academic coursework, and this means secondary vocational educators may be more interested in assessment for course improvement than assessment for certification or program evaluation. As compared with the

1980s, students are taking fewer vocational courses and there are fewer vocational teachers and fewer university programs training these teachers (Boesel & McFarland, 1994). Between 1982 and 1992, academic course-taking was up by twenty-two percent and vocational course-taking was down by seventeen percent (Vocational Education Journal, special pull out). Even with this trend, in 1992, almost all public high school graduates (97 percent) completed at least one vocational education course (Levesque et al., 1995, p. 7). Twenty-four percent of high school students were considered vocational concentrators completing at least three credits in a single vocational program area.

Although secondary vocational education is often associated with students planning to go to work after high school, most seniors plan to go on to some form of postsecondary education—49 percent plan to attend a four-year college or university and 22 percent a 2-year college or technical, vocational, or trade school (only fifteen percent of seniors plan to work full time; MPR Associates, 1995). Instead, students take single courses in order to learn more about a career or attain a specific skill related to work (i.e. word processing). Consequently, there is less emphasis on certifying mastery of employment skills among secondary vocational programs and more interest in assessments that are relevant to students taking only one or two courses, i.e., assessments that provide information to improve teaching and learning.

The enrollment pattern of vocational students at the postsecondary level is quite different than at the secondary level, and the assessment needs of postsecondary vocational educators are different, as well. In the late 1980s postsecondary vocational enrollments increased at the same pace as enrollments in general. Thirty-five percent of all undergraduate students were enrolled in postsecondary vocational education. In nonbaccalaureate programs, about one-half of these students reported majoring in a vocational area (Boesel and McFarland, 1994). Students in vocational courses vary in age, work experience and career aspiration. More important, postsecondary vocational students have varying motives for enrolling in vocational education courses. Some students enroll in a course to advance their career or begin retraining for a new career. This creates a need for assessments to improve learning and instruction, as was the case at the secondary level. Other students enroll in a sequence of courses in order to enter a particular career or to be certified for a particular job. In this case, vocational educators need to be able to certify that students have mastered relevant skills. In addition, accountability requirements apply to students who complete sequences of vocational courses, and staff must have assessment data to evaluate the success of these programs.

Knowledge and Skills

Historically, vocational education has existed to prepare students for specific jobs, but recently the nature of vocational education has changed. (Lazerson and Grubb 19xx). In the early 1900s, there was strong support from the business community for the federal government to fund vocational education in order to alleviate the scarcity of "skilled" workers through "skill" education. Businessmen alleged that the factory system had made the apprenticeship system obsolete and that it was now difficult and economically inefficient to allow informal, on-the-job learning in modern factories (Lazerson and Grub, 19xx). Hence, in 1917, the Smith-Hughes Act granted federal funds to public schools to develop vocational training.

Since 1917, the U.S. economy and workplace have broadened and diversified as have the goals for vocational education. The statement of purpose of the most current federal law, the 1990 Carl D. Perkins Vocational and Applied Technology Education Act, reads:

> "It is the purpose of this Act to make the United States more competitive in the world economy by developing more fully the academic and occupational skills of all segments of the population. This purpose will principally be achieved though concentrating resources on improving educational programs leading to academic and occupational skill competencies needed to work in a technologically advanced society (AVA Guide to the Act, 19xx, sec. 3)."

Our understanding of vocational skills has changed in two important ways. First, we have learned to place greater value on broad skills that relate to a family or cluster of jobs rather than narrow skills defined in terms of a single occupation. Second, we have discovered the importance of learning skills in context rather than in isolation, and schools are placing greater emphasis on learning in real world situations.

There are a number of ways to think about the skills and abilities that form the basis for vocational curriculum and assessment. Vocational educators in the United States traditionally decompose complicated occupational responsibilities into distinct, separable components and organize curriculum and instruction around them. For example, the trade of welding may be broken down into 50–100 distinct skills that are taught and practiced one at a time. This model of analyzing the demands of an occupation (called the job competency model) predominates in occupational training, occupational certification and licensing, and in the military (Wirt, 19xx). First, a job task inventory is constructed to identify the specific tasks that people will be expected to perform on the job. Second, this inventory becomes the basis for training and assessment. Typically, instructors or supervisors "check off" one by one those

tasks which a person can perform and indicate one by one at what skill level they can be performed. The Oklahoma assessment system is based on detailed task analyses of this type.

With a constantly changing workplace due to technological progress and international competition, the preparation individuals need for jobs or careers within different industries has changed. Often employers seek less job-specific training and more general workforce preparation training. Employers also seek individuals who can adapt to changing workplace conditions with a solid grounding in basic academic knowledge, the ability to handle responsibility, communicate and work with others, and solve problems. More and more secondary and postsecondary vocational programs are being asked to integrate academic and vocational knowledge and teach both broad and specific skills.

Skills can be thought of as lying on a continuum between general workforce preparation and specific occupational skills. Table 2 provides an example of skills from the health occupations at four points on this continuum. At its most general level, workforce preparation may be offered within traditional academic disciplines or broad industry areas that provide contextualized skills and knowledge. A second type of skills comprises more narrow industry- or occupation-specific skills intended to help individuals prepare for workforce entry. Occupational cluster and specific occupational skills describe a progressively more focused set of skills that a worker would need to master for a job within a group of related occupations or a specific occupational field. Although skills may transfer across industry areas, some are designed around specific workplace tasks that employees must routinely perform.

The type of knowledge and skills to be assessed can affect the choice of assessment methods. Traditional assessment forms, including multiple-choice and short-answer questions, are efficient ways to measure specific occupational skills. For example, a student enrolled in a word processing course may be asked only to master a set of very specific skills. However, as vocational educators focus on other parts of the continuum of knowledge and skills, these techniques become less effective. A student enrolled in a nursing program may need to master skills across the continuum. As will be explored in the next chapter, these skills are less well measured with selected-response tests. Federal law mandates greater integration of academic and vocational education and an emphasis on "all aspects of the industry," so attention to the more generic skills is a growing priority for vocational education. Table 3 shows the breakdown of our sample of assessment systems based on the type of knowledge and skills addressed.

Table 2

Continuum of Knowledge and Skills: Examples From the Health Industry

GENERAL WORKFORCE PREPARATION All Workers	INDUSTRY CORE SKILLS AND KNOWLEDGE Health Services	OCCUPATIONAL CLUSTER SKILLS Health Information Services	SPECIFIC OCCUPATIONAL SKILLS Health Information Technology
Read, write, perform mathematical operations, listen and speak	Be aware of the history of health Use health care terminology	Locate information in medical records Use computer programs to process client information	Evaluate medical records for completeness and accuracy Use a computer program to assign patients to a diagnosis-related grouping

Table 3

Knowledge and Skills Assessed in the Sample

	General Work-force Preparation	Industry Core Skills and Knowledge	Occupa-tional Skills	Specific Occupation Skills
California Testing Assessment Program (C-TAP)	✔	✔ (optional)	✔	✔ (optional)
Laborers-AGC environmental program				✔
Kentucky Instructional Results Information System (KIRIS)	✔			
National Board of Professional Teaching Standards (NBPTS)			✔	
Oklahoma Department of Vocational and Technical Education	✔			✔
Vocational Industrial Clubs of America (VICA)	✔			✔

Note: Oklahoma is phasing in assessments of occupational cluster skills.

Second, research has revealed the importance of context in learning. Cognitive scientists have begun to look at what people actually do in the work place (especially in "high performance" work environments) and have determined that often the knowledge of experts in a particular field is highly integrated and situational. Instead of viewing jobs as a list of skills or abilities, researchers are beginning to describe the actions of workers as performances in response to situations (Wirt, 19xx). In addition, experts gain and use information through working with others and creating shared knowledge to be used in the workplace.

This perspective leads to a different approach to instruction and assessment. Large units of performance become the focal point and the units are "situated" in a realistic context in which they might be encountered on the job. For example, students competing in a VICA competition would not be asked to list the steps to be performed when taking an order from a client; instead they would be asked to hold a realistic conversation with a person acting in the role of client and they would be judged whether they preferred all the desired behaviors. Similarly, the work samples students include in the C-TAP portfolios document performance of an occupational task in a real world setting. Vocational assessment are becoming more highly situated, and most of the assessments we sampled emphasized authentic performance of complex behaviors situated in a real world setting.

3. Types of Assessment

Interest in alternative assessment has grown rapidly during the 1990s, both as a response to dissatisfaction with multiple-choice tests and as an element in a systemic strategy to improve student outcomes. Alternative assessments range from written essays to hands-on performance tasks to cumulative portfolios of diverse work products. After a brief discussion of the strengths and weaknesses of traditional multiple-choice tests and other selected-response measures, the chapter describes four types of alternative assessments that might meet the needs of vocational educators and summarizes the major advantages of each type.

Comparing Selected-Response and Alternative Forms of Assessment

For decades selected-response tests have been the preferred technique for measuring student achievement, particularly in large-scale testing programs. Multiple-choice tests are one type of selected-response measure, a category which also includes true/false questions and matching items. In one form or another, selected-response measures have been used on a large-scale for 75 years. The defining feature of these measures is that respondents are given specific alternatives from which to choose, they do not have to create their response from scratch. This simplification leads to a highly efficient system of measurement. Students answer a large number of questions in a small amount of time. With the advent of optical mark sensors, responses can be scored and reported extremely quickly and inexpensively. Such tests provide an extremely efficient means of gathering information about a wide range of knowledge and skill. Psychometricians have developed an extensive theory of multiple-choice testing, and test developers have accumulated a wealth of practical expertise with this form of assessment.

Nevertheless, there are limitations to using multiple-choice and other selected-response measures. First, these traditional forms of assessment may not measure certain kinds of knowledge and skills effectively. For example, it is difficult to measure writing ability with a multiple-choice test. Similarly, a teacher using cooperative learning arrangements in a classroom may find that selected-response measures cannot address many of the learning outcomes that are part of the unit, including teamwork, strategic planning, and oral communication skills. In these cases, multiple-choice tests can only provide indirect measures of

the desired skills or abilities (e.g., knowledge of subject-verb agreement, capitalization and punctuation, and the ability to recognize errors in text may serve as surrogates for a direct writing task). Users of the test results must make an inference from the score to the desired domain of performance.

Second, when used in high-stakes assessment programs, multiple-choice tests can have adverse effects on curriculum and instruction. Many standardized multiple-choice tests are designed to provide information about specific academic skills and knowledge. When teachers focus on raising test scores, they may emphasize drill and practice and memorization in narrow ways without regard to the students' ability to transfer or integrate this knowledge. Instruction may focus on narrow content and skills instead of broader areas, such as critical thinking or problem solving skills (Miller & Legg, 1993). In addition, many think multiple-choice tests emphasize the wrong behaviors; few people are faced with multiple-choice situations in their home or work lives (Wiggens, 1989).

During the past few years, alternative assessment approaches have gained popularity as tools for classroom assessment and large-scale use. These alternatives are distinguished by the fact that students must construct responses rather than select from pre-specified alternatives. Constructions include written work as well as physical products and behaviors. In these case, students are asked to perform directly the desired behavior. Proponents of alternative forms of assessment believe they will alleviate some of the problems presented by multiple-choice tests. It is possible to measure a broader range of skills and ability using constructed-response approaches than selected-response measures. To measure writing ability, one asks students to write; to test oral communication, one has students give oral reports. In addition, alternative assessments permit the use of complex, realistic problems instead of narrow or decontextualized skills, which appear on many multiple-choice tests. Because of this, teaching to alternative assessments is desirable, because good test preparation will be good instruction.

Before describing the range of alternatives, we should note that alternative assessments are not without problems. In fact, they may have many of the same flaws cited for multiple-choice tests. For example, critics argue that poorly designed alternative assessments can also be very narrow, so that teaching to them may also be undesirable. In addition, alternative assessments have practical problems, including high cost, administrative complexity, low technical quality, and questionable legal defensibility (Mehrens, 1992). These flaws are of greatest concern for those assessments being used to certify individuals for work or to reward or sanction people or systems. These issues will be discussed in greater detail in Chapters 4 and 5. Table 4 lists some of the advantages of selected-response and constructed-response measures.

Table 4

Advantages of Selected- and Constructed-Response Measures

Selected-Response	Constructed-Response
Easier to develop, administer and score	Easier to incorporate real world settings
More efficient use of students' time	Better measure of real performance
Strong theoretical basis for judging quality of results	Appropriate for more complex tasks, critical thinking and problem solving
Familiar to teachers, students and community	Exemplifies more appropriate classroom practices
Good for measuring factual knowledge	

Table 5

Types of Assessment

WRITTEN ASSESSMENTS
Multiple-choice
Open-Ended
Essay, Problem-Based and Scenario
PERFORMANCE TASKS
SENIOR PROJECT: Research Paper, Project and Oral Presentation
PORTFOLIO

Types of Alternative Assessment

We use the term *alternative assessment* to refer to measures that require the respondent to construct an answer rather than choose from a pre-specified set of possible answers. Written items of this form, such as an essay or a solution to a mathematical problem, are often called constructed-response items. However, alternative assessments are not limited to written prompts or written responses. The label is commonly used to refer any form of work whose quality can be judged accurately, from live performances to accumulated work products.

There are a variety of ways to classify alternatives to multiple-choice tests (Hill and Larson, 1992; Herman, Aschbacher and Winters, 1992). In fact, since the range of constructed response types and situations is limitless and more formats are being developed all the time, it is unlikely there will be a single best system of categorization. For the purposes of this paper, we will use categories developed

by NCRVE that are clearly relevant to vocational educators (Rahn, et al., 1995). These categories distinguish four major types of assessment strategies: written assessments, performance tasks, senior projects and portfolios (see Table 5). For the sake of completeness, multiple-choice tests were included in the category of written assessments. We will maintain that distinction, although we will not devote much space to the multiple-choice alternative.

Written Assessments

Written assessments include activities in which the student selects or composes a response to a stimulus or prompt. In most cases the prompt or stimulus also includes printed materials (a brief question, a collection of historical documents, graphic or tabular material, or a combination of these). However, the stimulus can also be an object, an event, or an experience. Student responses to written assessments are usually produced "on demand," that is, the respondent does the writing at a specified time and has a fixed amount of time to complete the task. These constraints create greater standardization of testing conditions, which increases the comparability of results across students or groups, a theme which will be explored later.

Rahn, et al (1995) distinguish three types of written assessments. The first type is multiple-choice tests. As discussed above, multiple-choice tests are quite efficient, particularly for gathering information about knowledge of facts or the ability to perform specific operations (as in arithmetic). For example, in the Laborers-AGC program, factual knowledge of environmental hazards and required procedures is measured using multiple-choice tests. The Oklahoma testing program uses multiple-choice tests of occupational skills and knowledge derived from statewide job analyses. Multiple-choice tests can also be used to measure many kinds of higher-order thinking and problem solving skills, but they require considerable skill to develop.

Open-ended assessments are characterized by short written answers. The required answer might be a word or phrase (such as the name of a particular piece of equipment), a sentence or two (such as a description of the steps in a specific procedure), or an extended written response (such as an explanation of how to apply particular knowledge or skills to a situation). In the simplest case, short answer questions make very limited cognitive demands, asking students to produce specific knowledge or facts. In other cases, open ended assessments can be used to test more complex reasoning, such as logical thinking, interpretation or analysis. Scoring such questions is more complicated, as well, because the test developer must specify the desired response quite carefully and must develop a procedure for scoring partially correct answers.

The third type of written assessment includes essays, problem-based examinations and scenarios. These measures are like open ended assessment, except they extend the demands made on students to include more complex situations, more difficult reasoning, and higher levels of understanding. Essays are familiar to most educators; they involve a lengthy written response which can be scored in terms of content and/or conventions. Problem-based examinations include mathematical word problems and also more open-ended challenges based on real-life situations that require the student to apply knowledge and skills to new settings. For example, in Kentucky, groups of three or four twelfth grade students were given a problem about a Pep Club fund-raising sale, and they had to analyze the data, presenting it in graphical form, and make a recommendation about whether the event should be continued in the future. A scenario-based exam is similar, but the setting is described in greater detail and the problem may be less well-formed, calling for greater creativity. An example this type of assessment is the scenario portion of the C-TAP which requires students to write an essay evaluating a real life situation and proposing a solution (such as determining why a calf is sick and proposing a cure).

Performance Tasks

Performance tasks are hands-on activities that require students to demonstrate their ability to perform certain actions. This category of assessments covers an extremely wide range of behaviors, including designing products or experiments, gathering information, tabulating and analyzing data, interpreting results, and preparing reports or presentations. In the vocational context, performance tasks might include things such as diagnosing a patient's condition based on a case study, planning and preparing a nutritionally balanced meal for a vegetarian, or identifying problems with computer in an office and fixing them. Performance tasks are particularly attractive to vocational educators because they can be used to simulate real occupational settings and demands. Our cases included many example of performance tasks. For example, each Oklahoma vocational student had to complete two tasks designed and scored by his or her teachers. The VICA competitions are primarily life-like simulations, such as an emergency team responding to an accident victim.

Skill demands can vary considerably in performance tasks. Some tasks may demand that a student demonstrate his or her abilities in a straightforward way, much as they might have been practiced in class (e.g., adjusting the spark plug gap). One health trainee assessment involved changing hospital bed sheets while the bed is occupied, a skill that participants had practiced frequently. Other tasks can present situations that are novel and demand that a student figure out

how to apply learning in an unfamiliar context (e.g., figuring out what is causing an engine to run rough). Teachers participating in the NBPTS certification process must respond to unanticipated instructional challenges presented during a day-long series of assessment exercises.

As assessments become more open-ended and student responses become more complex, scoring becomes more difficult. A variety of methods have been developed to score complex student performances. The methods themselves range from simple to complex, depending on the sophistication of the components to be scored. In some cases, students are assessed directly on their performance, in other cases on a final product or oral presentation. For example, in the VICA Culinary Arts Contest, students prepare platters of cold food and a multi-course meal of cooked food using ingredients and equipment provided. Judges assess both the procedures used (by rating organizational skills, sanitation and safety) and the final product (by rating presentation and taste). Similarly, in the KIRIS interdisciplinary performance events students work together in groups on open-ended activities and then produce individual products. The group work is not judged, just the individual responses.

Traditionally, vocational educators have relied on performance-based assessment strategies to judge students' mastery of job-specific skills. For example, an automotive teachers judges whether a student can change the oil of a car by asking them to perform the task. However, other strategies may be required if that same teacher would like to assess her students' ability to understand the technical principles underlying an automotive engine.

Senior Project

Senior projects and portfolios are distinct from written assessments and performance tasks because they are cumulative, i.e., they reflect work done over an extended period time rather than work produced in response to a particular prompt or scenario. A senior project is conceived of as a culminating event in which students draw upon the skills they have developed over time. It has three components: a research paper, a product or activity, and an oral presentation all related to a single career-related theme or topic. The format is designed to be motivating, to permit involvement of people from business or community, and to encourage integration of academic and vocational ideas. For this reason, the process of implementing senior projects in a school often involves collaboration between teachers in many subjects who agree to guide the student's selection and accept the work for credit in more than one course.

All three components are organized around a single subject or theme, such as a traditional method of making furniture, the creation of an appealing store window display, or a fashion show. To complete the research paper, the student must conduct research about aspects of the subject they have not previously studied. The student draws upon library and other resources and produces a formal written paper. In the second stage, the student creates a product or conducts an activity relevant to the subject. This might include making something or doing community volunteer work for an extended period and documenting it. The purpose is to demonstrate knowledge or skills relevant to the subject. Finally, the student presents his or her work to a committee or public forum.

The length and complexity of the senior project make evaluation difficult. Schools that have implemented this type of assessment have spent a fair amount of time deciding how to judge the quality of the various elements. Their scoring guides reflect concerns about content, technical knowledge, organization and time management, extending knowledge outside of traditional school domains, communication skills, and even appearance (Rahn, et al., Module Four, p. U3-12). These are all subjective judgments, so great care must be taken to ensure that scores are accurate and meaningful.

Portfolio

A portfolio also is a cumulative assessment; it represents a collection of student work and a documentation of student performance. A senior project is a type of portfolio focused on a single theme. More generally, portfolios may contain any of the on-demand or cumulative assessments described above plus additional materials. Portfolios can contain a variety of products, including work samples, official records, and student written information. For example, in the C-TAP portfolio, students not only provide an artifact (or evidence of one if it is not portable) but they give a class presentation which is evaluated as part of their project. Records may include transcripts, certificates, grades, recommendations, resumes, and journals. Portfolios also often contain a letter of introduction to the reader from the student, explanation of why pieces were included, career development materials, letters from supervisors or employers, completed job applications, test results, and samples of work products. These may reflect academic accomplishment, industrial or career-related activities, personal skills and exhibits of accomplishments or performances.

Some portfolios are designed to represent students' best work, others are designed to show how work has evolved over time, and still others are

comprehensive repositories for all work. Both the KIRIS portfolios (for writing and mathematics) and the C-TAP portfolios (for a vocational area) are built around a selection of the student's best work. The C-TAP portfolio adds other types of assessment such as records (a resume) and a work sample artifact (a writing sample).

Portfolios present scoring problems because each is so diverse and because no two contain the same pieces. This variation makes it difficult to develop scoring criteria that can be applied consistently from one piece to the next or from one portfolio to the next. States that have begun to use portfolios on a large scale have had difficulty achieving acceptable quality in their scoring (Stecher and Herman, forthcoming), but they are making progress in this direction. One approach is to set guidelines for the contents of the portfolios so each contains similar components. Specific learner outcomes can be identified for each component and then techniques can be developed for assessing student performance in terms of these outcomes.

Table 6 shows the range of assessment types being used in the sites selected for this study.

Table 6

Types of Assessments in the Sample

	Written					
	MC	Open-ended	Essay, etc.	Perform-ance	Senior Project	Portfolio
California Testing Assessment Program (C-TAP)			✔	✔		✔
Laborers-AGC environmental training	✔			✔		
Kentucky Instructional Results Information System (KIRIS)			✔	✔		✔
National Board of Professional Teaching Standards (NBPTS)			✔	✔		✔
Oklahoma Department of Vocational and Technical Education	✔			✔		
Vocational Industrial Clubs of America (VICA)	✔			✔		

4. Criteria for Comparing Assessment Alternatives: Quality and Feasibility

As the last chapter suggests, vocational educators are likely to find more than one assessment strategy to serve a particular purpose. Two important criteria to use in selecting assessments for a particular situation are the quality of the information provided and the feasibility of the assessment process. This chapter describes these factors and compares selected- and constructed-response alternatives in terms of quality and feasibility.

Unfortunately, it is usually not possible to maximize both quality and feasibility, so vocational educators must strike a balance between them. As assessment becomes more authentic, it also becomes more expensive to develop, to administer, and to score. In addition, greater quality usually involves greater cost and greater commitment of time. There is no simple formula for balancing these factors. Ideally, educators would establish standards for quality based on the uses to which information will be put, and then allocate resources appropriate for meeting those quality standards. In addition, they would impose constraints based on practical needs that would not limit quality. In reality, this balancing act is more an art than a science, but we believe an understanding of the factors will lead to better decisions.

Quality of Assessments

The relative quality of the available alternatives should be a factor in selecting an assessment strategy. Concerns about quality are particularly important when assessments are used to make critical decisions, such as certifying individual skill mastery or rewarding successful training programs. Vocational educators face such decisions regularly, so it is important that they understand something about the technical quality of assessments.

The quality of an assessment can be judged in terms of three questions:

- How accurate is the information?
- How confident can we be in our conclusions about students or programs?
- Is the assessment fair to all students who take it?

These questions correspond to the psychometric concepts of reliability, validity and fairness. Given the present state of the art with regard to alternative assessments, not all approaches provide equally accurate information, support desired interpretations equally well, or provide all students with equivalent fair challenges. Table 7 summarizes some of the quality differences between selected- and constructed-response measures that are discussed below.

Table 7

Status of Selected- and Constructed-Response Measures with Respect to Quality of Information

Dimension of Quality	Selected-Response	Constructed-Response
Reliability	Automatic scoring is essentially error-free	Rating process can increase error
	Many items increase reliability of overall score	Fewer responses can reduce reliability of overall score
	Strong theoretical basis for measuring reliability	
Validity	Must make larger inferences from task to occupational behavior	Greater match between assessment task and real-world demands
		Variation in conditions can complicate interpretation of results
Fairness	Quantitative techniques help to identify potential unfairness	May have greater fairness because tasks are more authentic

Reliability

There are no perfect measuring tools, either in science, in the kitchen, or in education, so people who use tools to measure things need to know how much error there is likely to be in the information they receive. Reliability is a numerical index of the degree to which an individual measurement (such as blood pressure, volume of liquid, or a test score) is free from error. One common way to determine reliability is to repeat the measurement and see whether the results are the same. Thus, the carpenter's dictum, "measure twice, cut once." Another method to estimate reliability is to use a comparable tool and see whether the same result is obtained. When parents say, "she feels hot to me, let's get the thermometer and see if she has a fever," they are using a second method to confirm the results of the first.

The reliability of an assessment is the degree to which the score provided by the measure is accurate. If the test were administered again would students score the same, or is there so much uncertainty in the test itself that students would perform differently the next time? For example, in the case of VICA, if the contest was repeated would the contestants do equally well or would their rankings change? Another way to investigate reliability is to determine whether students would receive similar scores on another test of the same material or on comparable subparts of the existing test? These are the methods that are traditionally used to measure test reliability.

Most commercially available tests produce individual scores whose reliability is above 0.80. This means that 80% or more of the test score is due to "true" performance and less than 20% is due to measurement error. The acceptable standard may be higher (0.90 or more) when tests are used for important decisions. Commercial tests achieve these high levels of accuracy in part by obtaining many separate bits of information on each student. In an hour, a student might answer 50 or 60 multiple choice questions, providing a great deal of information about what he or she knows or can do. The Oklahoma assessments use selected-response items and their reliability is quite high.

For a number of reasons, alternative assessments may not be as accurate as multiple choice tests. Three features of alternative assessment have important effects in terms of reliability. First, the number of pieces of information obtained is usually far smaller than with multiple choice tests. Because the tasks are more complex and demand longer responses, alternative assessments produce fewer pieces of information about each student in a given amount of time. Students receive a handful of scores on the Kentucky portfolios that reflect many hours of work. Similarly, Kentucky performance events take a full class period and produce only one or two pieces of information about each participant. There are fewer data with which to make an overall assessment of student skill or ability and, as a result, the judgment may be less accurate.

Second, research in a number of fields has found that student performance on constructed-response measures varies more from task to task than on selected-response ones. As the demands of the task increase (in terms of complexity, breadth, integratedness, or any number of factors), consistency of performance declines. As a result, when alternative assessments are used not only are there fewer pieces of information about a student obtained in a given period of time, but the information may be less consistent. Consequently, the overall score is less reliable.

Third, scoring introduces additional errors not present with selected-response measures. Alternative assessments involve subjective judgments about the quality of complex student work. Rather than having an answer sheet scored with almost perfect accuracy by a machine, raters are asked to review essays, science projects, or pieces in a portfolio and assign scores on one or more dimensions. Both the CTAP and Kentucky portfolios require expert readers to review the material and assign scores using a general rubric. The same is true for the NBPTS assessment activities. The use of human judgments introduces an additional source of error. This reader inconsistency usually lowers the accuracy of final scores on alternative assessments. However, with practice readers can be trained to apply certain types of scoring rubrics with a high degree of consistency. Reports of inter-reader consistency above 0.80 are becoming more common. However, this value does not reflect the reliability of the final score, just the consistency of the rating process. The overall consistency of students' scores on alternative assessments is determined by the consistency of the raters and the consistency of the measure.

Validity

Scores can be accurate in the previous sense, but people can use them to draw the wrong inference. This is a problem of validity. For example, a student who knows how to solve mathematical word problems may perform poorly on a written test of word problems because of reading difficulties. The test score may be reliable (i.e., the student consistently makes mistakes on written word problems) but it would be incorrect to interpret the score to mean that the student did not know how to solve word problems in general.

An inference from a score that is justified is said to be valid. Whereas reliability is a feature of the measure, validity is a feature of the way the scores are interpreted by users. Consequently, assessments that are valid for one purpose may not be valid for another. For example, one might give a student studying to be a medical records clerk a multiple-choice test of spelling, grammar, and syntax to determine the student's ability to identify errors in textual material. However, this test might not provide a good measure of the student's ability to write grammatically correct information on a record.

One of the primary motivations for adopting alternative assessments is to increase the validity of the inferences by making the assessment tasks more like the real-world activities the tests are supposed to reflect. Constructed-response measures constrain the assessment to a rigid format, which can narrow the types of skills that are measured. Alternative assessments present students with tasks

that are more "authentic," i.e., they match more closely the activities performed in practice. As a result, it is hoped that the scores on the test will provide a better measure of the domain of interest than scores on a multiple-choice test. The Laborers-AGC environmental performance assessment duplicates conditions of the job and success on the assessment is thought to be highly predictive of success on the job.

There are a number of approaches to establishing the validity of an assessment for a particular purpose. One method is to have experts examine the assessment and judge whether its content is consistent with the domain it is supposed to measure. This is called *content validity*, and it is quite commonly used as the first step in building a case about the interpretation of assessment results. If the tasks to be performed are identical to tasks on the job, as they are in the case of Laborers-AGC, content validity may be adequate to satisfy the users of the information. The government is satisfied that mastery of the AGC performance tasks will produce competent workers. Similarly, the standards that underlay the NBPTS assessments were developed by committees of experts who reached consensus on the critical features of accomplished teaching. Extensive professional review forms the basis for the NBPTS standards, the appropriateness of the specific tasks, and the passing scores.

A second approach to validation involves comparing performance on the measure with current or future performance in a real-world setting. This is called *concurrent validity* or *predictive validity*, and it is based on the idea that a meaningful score will be positively related to real performance. Vocational educators in Oklahoma determined that scores on the multiple-choice tests were as good a predictor of future job performance as scores on lengthier scenarios, so they deleted the scenarios from the assessment program. This saved time and expense without reducing the validity of the scores for their intended purpose.

Construct validity is a third way of establishing the meaning of the scores on an assessment. This may be the appropriate technique to use when constructs being measured are complex and hard to define and when successful performance is a matter of judgment. Construct validation involves investigating the pattern of responses among a collection of assignments designed to measure similar and dissimilar concepts. In both of these cases it may be necessary to collect multiple sources of evidence to make a convincing case that the information is accurate for the intended purposes.

Alternative assessments present an additional validity challenge because they often have unstandardized components. The traditional model of testing controls both the form of the test and the procedures for administration so that everyone

has the same opportunity to perform and no one has access to special assistance. Similar standardization is possible for some constructed-response measures such as performance tasks, but other alternatives are inherently unstandardized. Variations in the content of the assessment or the conditions under which the assessment is administered make it more difficult to interpret the results. For example, senior projects and portfolios have built-in flexibility with respect to the conditions of performance and the content of the assessment. One student's C-TAP portfolio might contain different work artifacts and experiences than anothers'. Although it is possible to score both portfolios using a common rubric, there may be questions about the meaning of the two scores, since they were based on different products.

Fairness

Users of assessments must be concerned that irrelevant factors, such as family background or experience, might affect the scores of certain students. Assessments are unfair or "biased" if students who are otherwise equal with respect to the concept being measured perform differently on a particular question because of experience or knowledge not related to the underlying skill. It is not easy to detect possible bias. The most commonly used techniques involve careful review of measures by committees trained to be sensitive to factors that might affect particular groups of students. Expert reviews were used by NBPTS to ensure that the certification system was fair to teachers regardless of their population group or the socioeconomic status of their students. There are also complicated statistical procedures to determine if test items are biased, but the results of these procedures are often confusing. Researchers have found it difficult to understand what features of items lead to the differences they detect.

Many advocates of alternative assessments believe that these techniques are more fair to all groups because they involve more complete tasks and permit students to address the tasks in ways that are meaningful to them. However, there has been very little rigorous research on the fairness of alternative assessments. If vocational educators are going to use assessments with students from diverse backgrounds they need to be sensitive to potential unfairness in the measures they select.

Feasibility of Assessments

Practical considerations will also play an important role in choosing among assessment alternatives. In general, alternative assessments are more difficult to

develop, more time-consuming to administer, more troublesome to score, and they yield results that are more difficult to explain than selected response tests. In purely practical terms, selected-response tests are a model of efficiency. Potential users of alternative assessments need to be concerned about feasibility issues such as cost, time commitments, complexity, and acceptability to key stakeholders. These features are discussed below and summarized in Table 8.

Table 8

Feasibility of Selected- and Constructed-Response Measures

Dimension of Feasibility	Selected-Response	Constructed-Response
Cost	Relatively inexpensive to develop, administer and score	More expensive to develop, administer and score Teachers benefit from participation in scoring
Time	Efficient use of class time Few demands on teacher preparation time	Consumes additional class time Teachers require more preparation time Embedded-tasks may not detract from class time
Complexity	Relatively easy for developers and users.	May require special skills to develop May need special materials to administer Difficult judgments make scoring difficult
Acceptability	Familiar and well-known Higher reliability leads to greater confidence	Growing popularity among educators Unfamiliar to community members Credibility with employers

Cost

In general, alternative assessments are more expensive to develop, administer and score than selected-response tests (Hoover, 1995; Stecher, 1995). For example, because the tasks themselves are more complex and contextualized than multiple-choice items, they take more time to draft, pilot, and revise. In addition, many of the costs increase as the assessment becomes less constrained and more "authentic."

However, most potential users are not overly concerned about test development costs. They are concerned about the demands of test administration and the costs of scoring. Scoring alternative assessments can be many times more

expensive than scoring selected response tests. It costs pennies per student per class period to score multiple-choice test and produce detailed score reports. By contrast, it costs dollars per class period per student to score essays, performance tasks and portfolios. For example, Stecher and Klein (in press) found that science performance tasks cost 30 times as much as multiple choice tests per class period and 100 times as much for an equally reliable score. Commercial publishers who provide writing assessments charge about $5 per student for scoring essays and reporting a single score, either holistic or analytic. The NBPTS costs remain extremely high, in part, because of the complexity of judging candidate performance.

On the other hand, the use of alternative assessments may bring unanticipated benefits that offset some of the additional costs. Teachers report that scoring performance assessments is an effective staff development activity. The process of reviewing student work and evaluating it with respect to standards helps teachers develop better appreciation for the range of student performance, weaknesses in some students' presentation, common misconceptions and problems encountered by students, the alignment between curriculum and assessment, and other features that relate to instructional planning and may improve teaching and learning. If scores provide more valid indicators of job preparation, then they may be worth some added cost.

Time

In addition to those costs that must be borne directly, alternative assessment place greater time demands on administrators, teachers and students. For example, alternative assessments usually require more class time to administer than multiple choice tests. The use of class time for assessment can have negative consequences on instruction. Scoring also commands a great deal of time. There are advantages to having teachers score their own students' work. For example, they learn more about student performance, and there is no added cost for hiring outside scorers. However, scoring is an extremely time-consuming task, and teachers should be aware of the demands scoring may place on their preparation time.

When assessments are embedded in classroom instruction, such as senior projects and portfolios, the distinction between assessment time and learning time is blurred, and the time problem is less troublesome. This is the case with C-TAP and with the KIRIS portfolios. These assessments do not place the same significant additional demands on classroom time as do stand-alone performance assessments.

Complexity

Alternative assessments are more complex than traditional tests in a number of ways, including the situations that prompt student responses, the kinds of materials that are involved, the scope of the tasks, the cognitive demands placed on students, the procedures for collecting responses, and the procedures for scoring. As noted above, it is partly this complexity that makes alternative assessments more difficult to develop, administer and score, which increases their cost. The complexity also demands more sophistication on the part of users. For example, it can be more complicated to administer performance assessments that involve equipment and materials than to administer pencil and paper tests. In the case of Laborers-AGC, the tasks can include the use of heavy machinery, hazardous materials and dangerous working conditions. The equipment makes administration more complex and places greater demands on task administrators, who need to be specially trained to work under these circumstances. Similarly, it may take greater expertise to develop good portfolio tasks, to devise scoring rubrics for senior projects, etc., than to administer and score selected-response tests. The additional complexity inherent in alternative assessments may create practical problems for some educators and some educational settings. Additional training may be required, as well as additional equipment and materials, storage space, and facilities for assessment.

Acceptability

To have any practical value, assessments must provide information that is credible to the people who will use the results. In the case of vocational assessment, this includes the usual educational audiences, including students, teachers and program directors, but it also includes potential employers, labor leaders and other community members. If an assessment fails to meet reasonable technical standards, its credibility may decline in the eyes of some audiences. For example, Kentucky teachers still have doubts about the appropriateness of KIRIS as an accountability tool. But even if the assessment is found to be reliable and valid, people who are only familiar with traditional tests may not place much trust in scores generated by performance tasks, senior projects, or portfolios. Part of this discomfort may be due to unfamiliarity, and it should be possible to overcome this problem with training. On the other hand, one of the advantages of alternative assessments is that employers and other stakeholders may give greater credibility to scores based on authentic performances than to traditional test results. It appears that this has been the case for the Laborers-AGC environmental program and for the NBPTS certification program. It is true for VICA, as well.

5. Other Issues in Assessment Planning

Quality and feasibility are important factors in assessment planning, but they do not always present themselves in the general ways discussed in the previous chapter. The case studies produced examples of other administrative considerations related to quality and feasibility that also can affect assessment planning. We identified six additional issues that may confront the developers of assessments. Not all programs will need to address all these issues, but as a set they illustrate the additional complexities that may arise in assessment planning. The six considerations are:

(1) Single or multiple measures?

(2) Consequences of performance?

(3) Embedded or stand-alone tasks?

(4) Degree of standardization?

(5) Single or multiple purposes?

(6) Voluntary or mandatory participation?

Single or Multiple Measures?

There are obvious advantages in terms of efficiency for basing an assessment on a single measure, but there are reasons to prefer multiple measures, as well. We saw both options in our case studies. Although the Oklahoma assessment program contains both performance assessments and standardized multiple choice tests, Oklahoma relies on the multiple-choice test to determine whether individual students have mastered the curriculum in each vocational area. Each entrant in a particular VICA competition completes just one occupational task. It might be argued that the C-TAP portfolio is a single measure, but, in reality, it subsumes many measures. The C-TAP portfolios can contain other kinds of assessment results, such as test scores or competitive awards.

The principal advantage of single measures is efficiency (see Table 9.) Oklahoma provides a good example of this. In the past, the Oklahoma state vocational testing program had two elements: multiple-choice items and realistic scenarios followed by sets of related questions. The scenarios were more complex to develop and score, and the SDE decided that the multiple-choice items did an adequate job of predicting job-related performance. As a result, they opted to

Table 9

Advantages Associated with Single and Multiple Measures

Single Measure	Multiple Measures
Efficiency of planning, administration and scoring	Includes different types of skills and abilities
Reduced time and cost	Greater confidence in interpretation of student performance (i.e., greater validity)
	Drive programs toward more diverse curriculum and instruction

drop the scenarios from the testing program, and they achieved some reduction in cost and resource demands, as a result. Similarly, VICA made a determination that a single performance event was adequate for a competition whose goals are primarily honorary. Even with this simplification, they find that it is quite difficult to prepare the task specifications and scoring guides and train the raters for a single activity per occupation. Multiple activities would be prohibitive in terms of time and resources.

Educational researchers recommend the use of multiple measures primarily for reasons of validity that come from having alternative windows on behavior. The National Board strongly believes in multiple measures, arguing that the job of teaching cannot be captured in a single type of assessment. Laborers-AGC uses both a multiple choice test of knowledge and a performance test of ability to perform essential job tasks. Particularly as it relates to health and safety issues, the government dictates that candidates must demonstrate both job-related knowledge and the ability to perform essential tasks. KIRIS combines three types of student achievement measures—open-ended individual tasks, group performance events and individual portfolios—as well as noncognitive measures (attendance, retention, etc.) into a single school accountability index. They believe this provides a more complete picture of the multiple outcomes of schooling. The added quality comes at a price, because multiple measures are more time-consuming to develop, administer and score.

A second advantage is that multiple measures suggest more varied types of instruction and preparation. In high-stakes situations, single measures can lead to undesirable narrowing of instructional content and strategies (Shepard, Koretz). Under the same conditions, it is better to send richer signals to teachers and to force them to prepare students to succeed in many assessment situations, including forced-choice, open-ended written, performance, and exhibitions.

Consequences of Performance

Many aspects of the assessment are affected by the consequences attached to the use of the assessment results. The stakes—that is, the degree to which the outcome is associated with important rewards or penalties—can affect the character of the assessment, its credibility, the validity of scores, and the influence it has on instructions (Table 10). Assessment may have high stakes for individuals or for programs or schools. For example, a person will be denied certain jobs in the environmental hazard industry if he or she fails to pass the relevant Laborers-AGC examination. The National Board hopes that teachers who pass its certification assessment will earn respect, position and eventually greater rewards because of their proven skills. KIRIS, on the other hand, has no consequences for individual students but serious consequences for schools. Continued high performance may lead to financial rewards for teachers, and continued low performance can lead to intervention by the state Department of Education.

Table 10

Advantages Associated with Low and High Consequences

Low Stakes Assessments	High Stakes Assessments
Less pressure to "teach to the test" (possibly narrowing curriculum and instructional approaches)	Greater motivation to perform well on the assessment
More cooperative rather than competitive atmosphere	Greater emphasis on teaching the skills being assessed
Lower cost to develop and score	
Lower demands for reliability and validity	

High stakes have two major effects: they increase the level of scrutiny placed on results and they influence people's behaviors in anticipation of the assessment. A licensing examination is a good example of the latter situation, and the National Board comes closest to that model in the cases we studied. Certification carries with it valued consequences—in at least one state, Board Certified teachers receive a salary bonus. Because the certification results in a valued outcome, teachers must have confidence in the process. There have been many instances in which both licensing and employment assessments have been challenged in court by people who failed to pass and therefore were denied a benefit. For this reason, technical quality is an essential element of the assessment.

The consequence of the premium on technical quality is that more time must be devoted to development and more research put into measuring reliability and validity. As a result, high-stakes assessment can be far more costly than assessment used for low stakes purposes. KIRIS models some of these conditions. Since rewards are offered based on improvements in school accountability scores, the Kentucky Department of Education must ensure the quality of the scores. This necessitates additional research and development with their associated costs. In comparison, Oklahoma is committed to multiple-choice testing, in part, because such tests can produce reliable scores more efficiently than the more complex alternatives they tried in the past.

The second effect of high stakes is that changes in people's behavior may also affect the meaning of assessment results. In the case of VICA, individual levels of interest and anxiety affect performance, and scores may not reflect what would be anticipated under normal circumstances. Stakes may also drive teachers to unusual behaviors, both desired and undesired. In the case of KIRIS, researchers have detected both positive changes in curriculum emphasis and negative increases in inappropriate test preparation practices (Koretz, et al., in press).

Embedded or Stand-Alone Tasks?

Traditionally, tests are distinct events that follow, but are not part of, an ongoing learning process. Assessments are administered at the culmination of a set of learning activities. For example, when Laborers-AGC environmental trainees complete a safety unit each must demonstrate mastery of that unit by passing a performance test. Similarly, the VICA skills competitions occur independent of any classroom training.

There are alternative models in which assessment events are built into instructional activities as part of the curriculum or in which the products from meaningful learning activities are gleaned for the purposes of assessment. We use the term curriculum-embedded assessment to refer to both situations. C-TAP and the portfolio component of KIRIS are the best example of embedded assessments in the group we studied. As students complete work internships they capture evidence of their experience and include these in their C-TAP portfolio. Similarly, Kentucky students select their best classroom products in writing and mathematics to include in their portfolios.

Such stand-alone assessments have both logical and practical advantages. (See Table 11.) Stand-alone assessments serve as markers for accumulated knowledge

Table 11

Advantages Associated with Instructional Integration

Stand-Alone	Curriculum-Embedded
Greater flexibility for designing assessments	Greater efficiency
Greater standardization across classrooms	Greater authenticity with respect to the classroom lessons
Greater simplicity of administration	
Greater impact as cumulative or "capstone" event	

and skills. This approach gives the assessment developer greater flexibility to design tasks without worrying about the specific instructional activities employed by each teacher, which simplifies the design and administration of assessments.

Embedded assessments have advantages as well. First, they may be more efficient, not requiring teachers to set aside valuable class time for testing. Second, they lead to judgments based on student products from less artificial conditions. However, the conditions of performance are different for every classroom, it is difficult to interpret comparisons based on embedded assessments, and we know of no operational assessment system that relies entirely on such measures.

Degree of Standardization

Most state testing programs are examples of standardized assessment systems, i.e., assessment conditions are the same in every location. Individual sites have little flexibility to change what is assessed or how it is measured. For example, the Oklahoma Department of Education develops and maintains the vocational testing program, and each institution implements the tests according to standardized procedures. Similarly, the Laborers-AGC assessment is planned centrally and administered in the same fashion in every site. NBPTS, KIRIS and VICA are also centralized systems, but local teachers and programs have a degree of influence on selected aspects of these assessments. Kentucky teachers select the assignments that generate student work for the portfolio component of the assessments, NBPTS applicants provide materials drawn from their teaching experience, and active local VICA chapters can contribute to the planning of the national competitions. C-TAP, by comparison, is far more adaptable. Teachers adapt the portfolio framework to reflect their local emphases.

The advantages of a standardized approach include consistency of implementation and comparability of scores. (See Table 12.) All teachers administer the assessment according to the same rules so results can be compared from one site to another. For example, comparable tests are given in each Oklahoma vocational program, and students who pass the test in one school are demonstrating mastery of the same materials as those who pass the tests in another. Similarly, students take the same constructed response tests and performance events throughout the state of Kentucky. All VICA contestants perform the same job-related tasks, as do all candidates for Laborers-AGC certification.

Table 12

Advantages Associated with Standardization

Standardized	Adaptable
	Greater sense of ownership among teachers and students
Greater consistency of implementation	More relevant to local curriculum and community
Greater comparability of results across sites	More meaningful to individual students

Adaptability has advantages, as well. Most notably it permits assessment to be more responsive to local needs. For example, teachers can customize C-TAP portfolios to the curriculum emphasis in their course and the employer base in the neighborhood in which they are located. Students who use the C-TAP portfolios in a health program include work samples related to health, while those in a transportation program assemble different kinds of work samples. Permitting individual teachers to tailor the assessment to their local needs has other positive effects. For example, teachers may endorse the assessment more because it can be made more relevant to their programs. This is an important rationale for using portfolios in the Vermont assessment program (Koretz, et al, 1991). The most familiar form of adaptable assessment is classroom testing. Teachers are responsible for designing their own tests, and they implement them to meet their own classroom needs.

It is possible to combine adaptable and standardized components, as is done in KIRIS. The on-demand components—open-ended questions and performance events—are the same everywhere, while the portfolios differ from class to class based on the tasks assigned by the local teachers. Similarly, the National Board certification process has some flexible elements along with some standardized ones. Candidates for National Board certification supply a videotape of their

own lesson and do an analysis of their own instructional planning and decision-making. The unique individual video elements are combined with common assessment center exercises so NBPTS obtains a profile with both unique and shared elements.

Single or Multiple Purposes?

Most of the assessment systems we reviewed were designed to serve one of the three purposes described previously: providing information for instructional improvement, certifying student mastery, or evaluating program success. For example, the Laborers-AGC and NBPTS examinations are designed specifically to measure mastery of job related knowledge and skills, rather than diagnosing skill deficiencies or evaluating program effectiveness.

However, at least one of our cases, Oklahoma, involved an assessment system developed with multiple uses in mind. The Oklahoma system is supposed to serve dual purposes. Students' scores are aggregated to the program level, where they are used by the state to monitor program effectiveness and contribute to funding decisions. In addition, scores are reported to teachers, who use the scores to identify weaknesses in their curriculum or instruction and make adjustments.

Although it is easy to differentiate the three purposes of assessment in the abstract, in reality they are interrelated in many ways. Assessment results do not necessarily support only one use. For example, if too many students failed the Laborers-AGC examination it might suggest problems with the instructional program, and if all students passed the examination it would suggest that the instructional program was effective. Nevertheless, information gathered with one purpose is mind is likely to be better suited to that use than to another.

Some of the advantages of single purpose assessment are summarized in Table 13. A major advantage of a single purpose assessment is that it can be made as relevant as possible to the needs of the users. This can lead to efficiencies in design, administration and reporting. For example, when designing an assessment for program evaluation it is possible to sample students and tasks rather than having all students complete all exercises. Sampling reduces the burden on participants while still providing trustworthy aggregate information for judging the overall effectiveness of the program. However, this approach would not be appropriate for determining individual mastery because each student does not respond to enough items to provide a valid score.

Table 13

Advantages Associated with Single and Multiple Purposes

Single Purpose	Multiple Purposes
Greater clarity in design and reporting of information	Greater efficiency in use of assessment resources
Less conflict between competing demands	Greater alignment among users of common data
Shorter and more focused assessments	

In theory, it is possible to design multi-purpose assessments, but this goal has been difficult to achieve in practice. One problem is that the size of the assessment increases as the number of purposes increases. Another problem is that different purposes can lead to conflicting demands. In Oklahoma, the use of the assessment for program accountability and student learning are complementary, but there are some tensions between them that have to be resolved. For example, although Oklahoma provides common curriculum handbooks, not all teachers use them. Therefore, some teachers and students do not view the test as complementary to their curriculum and they fail to see the utility in taking a test that is only useful for state reporting purposes.

Similarly, an assessment designed to provide individual diagnostic information needs to produce scores at a finer level of aggregation than a test which does not have to help students and teachers plan instruction. For example, one might need to know whether students have learned specific grammatical conventions as a basis for instructional planning—should the class review the use of apostrophes in the possessive form? However, in a mastery setting, it probably is adequate to sample a variety of grammatical conventions within a written communication task.

Voluntary or Mandatory Participation?

One important element in the assessments conducted by Laborers-AGC, VICA, and the National Board is that participation in the program is voluntary. As a result, the individuals who sit for these tests do so by choice. The alternative is to require participation by everyone, as is done in state testing programs such as KIRIS and Oklahoma.

Table 14 illustrates some of the advantages associated with voluntary and mandatory participation. Students who are participating in a program

Table 14

Advantages Associated with Voluntary or Mandatory Participation

Voluntary	Mandatory
Greater commitment, hence more optimum performance (validity)	Value of assessment accrues to everyone
Strong influence on curriculum and instruction	Greater comparability across units
Greater learning from participation in the assessment	Strong influence on curriculum and instruction

voluntarily are often more motivated to do well, because they have made a commitment to the outcome. If, in their desire to be successful, they pay more attention to the tasks, focus their energies, and make more efficient use of time, it may even increase the validity of the assessment results. Voluntary participation may also increase the value of the assessment as a signaling tool because students and staff attend to it more. Teachers may adjust their curriculum based on scores and students may change their study habits. The assessment may have greater utility as a lever for reform because it is given greater credence by program participants. Increased attention may also enhance the educational value of the assessment experience itself. Finally, those who choose to participate often are more engaged in the learning experience than those whose participation is compelled.

There are advantages to requiring participation, as well. Oklahoma, KIRIS and C-TAP can be motivating because they are required. Similarly, teachers may attend to the content of the tests since participation is mandatory, although the degree of influence may be affected more by the consequences than by the level of participation (see above). Required assessments affect all participants, so whatever value is obtained accrues to everyone, not just the self-selected few. In addition, it is more likely that the assessment result will be useful for comparisons across program units when all students participate.

In most instances program developers have little choice over this aspect of assessment. The program context dictates whether all participants must take the assessment. However, there are cases in which design decisions can affect the amount of testing required of individuals and the use of the scores, and therefore indirectly the emotional and psychological aspects of participation. For example, some state testing programs report scores on every student, which necessitates that every student complete the full test. Other states report only aggregate scores (e.g., at the classroom, school or district level) which permits them to use

matrix or item sampling. While all students must participate, each takes far fewer items and some of the negative associations that accompany extended testing programs are lessened. In some instances, e.g., in Kentucky, a sample of students is selected to participate in the performance events, reducing further the perceived burden and giving participation an aura of specialness.

6. Discussion

This project was undertaken to investigate the utility of alternative assessments for vocational education and to provide vocational educators with guidance in evaluating different strategies for assessment. The case studies provide a rich set of illustrations of the range of constructed-response measures that are available to vocational educators and the purposes they might serve. However, the cases do not identify a set of best practices or a simple formula for choosing among alternatives. Instead, this research suggests that vocational educators will have to make their own choices from a growing ranges of options. That information may be interpreted as good or bad news. Some may long for a simple all-purpose solution; for them, the results of this study will be disappointing. Others may be excited to learn that they have considerable freedom to craft assessment systems to meet their needs.

This project will have value if it helps vocational educators make better assessment choices. To that end, we discussed a number of elements that need to be factored into assessment decisions. The choice of an assessment strategy should depend on the purposes to be served, the quality of the information desired, and the feasibility of different alternatives within the local context. Our six cases illustrated how different programs have crafted assessment systems to meet their specific needs. In addition, the case studies suggested a number of other factors educators must address when thinking about assessment systems.

In the next two sections we will try to illustrate how the information from the previous chapters might be used to address the needs of vocational educators. To do this we have chosen two common situations--assessment for program improvement and assessment for certifying student mastery--and we will illustrate how assessment planning can be informed by the results of this study.

Example: Developing Assessments For Program Improvement[1]

Dale McIver's Problem

Dale McIver teaches Office Automation at Watson Tech, an Area Vocational Technical School in Dade County, Florida. Dale teaches a three course sequence leading to a certificate in Office Machine Operation, but few of her students

[1] The individuals, schools and programs in this example and the one that follows are fictitious.

complete the full sequence. Instead, her classes are primarily composed of students wanting to gain some initial familiarity with computerized text and data processing or wanting to upgrade their skills in particular ways. Dale is frustrated with her current grading system, which is based on unit tests from the class workbooks. The test scores do not give her or the students enough information about the students' abilities to respond to realistic office demands that involve automated equipment. She is looking for an assessment system that would be more engaging for her students, help them understand their own strengths and weaknesses, and provide her with information to improve her instruction. She believes there is too much emphasis on rote learning of commands and functions. Instead, she wants her students to be better problem solvers when it comes to using computers in the office environment.

Developing a Solution to Dale's Problem

Dale's situation should be familiar to vocational educators, because the changes she is experiencing are widespread. In fact, part of the motivation for this project was to address these new challenges facing vocational educators. Our approach to solving Dale's problem is to review the elements of the situation in the order described above: purposes, quality, and feasibility.

In this situation the broad purpose for the assessment is unambiguous. Dale's interest is improving the courses she teaches. She wants information to help students focus their efforts and help her determine which skills need additional emphasis. She also recognizes there have been changes in the nature of the skills to be taught and the needs of the students who are enrolling, and hopes the assessments will be responsive to these conditions. In particular, she wants information about how well students would respond to realistic problems.

All of the alternative assessment methods described above could be used in this situation. Students could write extended descriptions of the procedures they would use in a particular office situation. Dale could develop realistic office tasks students must perform and then judge the products they produce. A course-long project culminating in a formal presentation is a possibility, but it seems less well-matched than the other alternatives. A portfolio containing a collection of tasks would work well. If part of Dale's goal is for students to build a repertoire of solution strategies and a command of the technology, then a collection of successful products might contribute to this development.

It is important to recall Dale's purpose and to consider how well each type of assessment would help the students and Dale to improve. We cannot analyze the situation completely without a clearer understanding of the nature of the

"problem solving" skills Dale hopes to foster and how well they could be embodied in the assessment. For example, does she expect students to be able to revise a document based on editors marks, produce a presentation quality organizational chart based on a sketch, or compile a report that includes text, tables and a graph? Does she expect students to find a file without knowing its filename, recover a document after a power outage, or repair a faulty disk drive? These are all problems one might encounter in an office setting, but they require different knowledge and skill to solve. Some might be measured adequately with written questions, others with performance tasks, and others in the form of an extended project.

Dale's plans for grading also play a role in selecting the assessment. Does the manner in which students produced the product matter or just the quality of the result? Some grading standards may be more helpful than others, just as certain types of feedback may be more informative than others. Dale would know (or would be able to find out) the answers to these questions, so she could factor them into her evaluation of the alternatives.

One of Dale's purposes for the assessment is program improvement. To improve her course it is essential that the assessment provide information that is easily linked to instruction (either to particular units or to behaviors). Will Dale know what to do if students do poorly on one aspect of the assessment?

In Dale's case, it is not necessary to place a premium on technical quality. Students will have multiple opportunities to perform in class, and the assessment results will not be used for critical decisions. Dale need not be overly concerned about the accuracy of scoring. Teachers make judgments about student performance all the time, and there is no reason to think Dale will be unable to judge fairly the responses to the assessment tasks. If the results are going to be used to help students judge their own skills and to inform changes in lessons, there is no reason to worry about validity. However, if Dale wants to draw inferences about broader behavior, such as "problem solving in the office environment," she would need to collect far more information to test the validity of that inference.

Although all the alternatives are feasible, each would increase the burden on Dale's time compared to her current assessment methods. The options do not involve great financial costs, but Dale would have to be willing to bear added preparation burdens. The portfolio appears to be the most demanding because of the time to organize it and assess the products at the end of the term. None of the alternatives appears to be so complex that Dale would need specialized help or so unusual that they would encounter resistance from students or faculty.

Administrative issues are less germane to this situation. The vignette does not clarify whether Dale's problem is unique to her or whether it is shared by other teachers in the district. It might be possible to share the development effort with other Office Automation teachers, so long as all remained engaged and remained committed to the work.

Example: Developing Assessments for Certification

J. C. San Martino's Problem

J. C. San Martino is the Coordinator of the Automotive Repair program for the Fort Meede School District. He supervises seven teachers in five high schools and one vocational school. Historically the program has been very successful in preparing students for entry-level jobs in service stations and repair shops. In the past two or three years local employers in various parts of the district have complained that graduates were not as good as they used to be. It has been hard for J.C. to respond because the complaints are all different, but the common thread seems to be that students are very good at some things but have gaps in their training. Employers are losing confidence in the district's program, and they are growing cautious about hiring graduates. Although every school uses the same curriculum and teaches to the same set of competencies, each instructor is responsible for his or her own testing and grading. J.C. thinks that a common assessment system might help him raise standards in the program, assure employers that graduates are competent, and encourage instructors to provide more consistent training. He has considered offering employers a guarantee that graduate from the auto repair program will meet agreed upon standards, but is not certain he could develop a system that would support this claim. He also wants to be sure that the assessment system provides information to help teachers improve their programs.

Developing a Solution to J.C.'s Problem

Comments from employers have led J.C. to question the quality and consistency of the training being provided in the district. He wants to use assessment to certify student competence, but he realizes that he must also provide information for program improvement or he will never achieve the certification goal. The problem is complicated by the fact that there are five different institutions, and there appear to be problems with individual schools as well as with the overall program. So, the assessment system needs to provide information to certify mastery and to identify shortcoming in the instructional program at each location.

There is more than one type of assessment that would address J.C.'s concerns, but not all the approaches we described in Chapter 3 are equally helpful. Because the auto repair program is organized around a set of competencies and because employers have expressed concerns about specific skills, the assessment should relate to these specific program elements. Written tests and performance tasks can both be useful in this regard. Senior projects, are less helpful for providing information about specific learning outcomes. Portfolios might be constructed in such a manner that they contain information linked to core competencies and skills.

One issue J.C. must address before designing the assessment system is whether the course curriculum is aligned with the needs of local employers. It might be that the demands of entry level auto repair jobs have changed in the past few years and the curriculum itself needs to be updated. Students may need to acquire new skills, such as using new finishes or new application procedures, or they may need to be prepared to work in different arrangements, such as multi-person teams. Reaffirming the link between course content and employer needs is an important first step.

There are many ways to measure the competencies students are being asked to master, and J.C.'s biggest challenge will be choosing methods that strike the right balance between quality and feasibility. Strictly speaking the "guarantee" he hopes to give to employers is not a binding contract, and there is no reason to apply the same quality standards one would apply to a licensing examination designed to protect public health and safety. Nevertheless, J.C. wants the effort to have merit, and he is particularly concerned about the consistency of performance across schools. It is not enough for each instructor to check off each competency as it is mastered, J.C. wants to impose some common external measures that have credibility with employers. Therefore, he may want to use common written exams to test some knowledge and common performance tasks to measure some applied skills. Using the same scoring procedures for all schools will provide the comparability he desires.

On the other hand, standardized measures (either written or performance) require time and effort, and too much testing takes away from learning time and may annoy participants. Because there are many ways to measure the competencies, it may make sense to measure some skills using quick and less intrusive methods (such as checklist initialed by the instructor) and measure other more difficult or important skills using more time-consuming methods (such as standardized written tests or standardized performance assessments). In some cases, improvement may come from merely requiring students and instructors to monitor their progress against a master list of competencies. A portfolio in which students compile evidence of mastery of all course competencies is an alternative he might consider.

It probably would be wise for J.C. to involve all the auto repair instructors in the development of the assessment system. This will help them understand the need for the system and increase their commitment to using it. More importantly, the instructors will have useful insights into ways to measure various skills. For example, it may be possible to have embedded components, in which existing class projects become the measurement tool for certain skills. Teachers also may have a better sense of the demands that certain choices will place on their time and the students' time. If the system is meaningful to instructors it will increase the likelihood that they will participate enthusiastically. If instructors incorporate performance on the assessment into course grades, the system will have more meaning for students, as well.

Conclusions

The results of this study indicate that alternative assessments can be useful tools for vocational education, but vocational educators must learn to be wise consumers with respect to assessment. Our examination of cases illustrates the breadth of assessment options that are available, from open-ended written assessments to performance tasks to portfolios. Each has been used effectively on a large-scale in at least one location, and all appear to have potential for vocational education.

For vocational educators who do not know where to begin, we have suggested an approach to choosing among alternative assessments. The first step is to clarify the purposes of the assessment and the specific conditions of the vocational context. These conditions might include the needs of constituents, demands for accountability, and the nature of the skills to be assessed. Next, one should consider a wide range of assessment options. The case studies illustrate a few alternatives that have been used in practice. Reading the complete descriptions in the Appendices conveys a fuller picture of demands of each situation and the strengths and weaknesses of the approaches taken. A thoughtful educator would supplement our reports with information from colleagues and professional organizations about other assessment methods that might be used or adapted.

Educators must also consider the issues of quality and feasibility. The manner in which the information will be used determines the level of technical quality that needs to be achieved. In general, the more importance that is attached to the use of the information, the higher the levels of reliability and validity that are appropriate. However, quality concerns should be balanced against practical realities. Cost, time burdens, and acceptability by stakeholders are also important considerations in selecting assessment methods. Some approaches are cheaper, less intrusive, and more familiar than others.

In the end, there is no single assessment approach that is best for all situations. However, there are fairly simple considerations that can guide assessment planning. This study illustrates the breadth of alternative approaches that exist, their utility in the vocational context, and some procedures vocational educators can use for making choices among them.

Appendix A
Kentucky Instructional Results Information System (KIRIS)

The Kentucky Instructional Results Information System (KIRIS) is a multidimensional measurement and assessment system that supports the statewide educational accountability system in Kentucky. It was initiated by the Kentucky Department of Education in 1991 in response to a comprehensive statewide educational reform law. KIRIS collects data on cognitive outcomes in grades 4, 8, and 12, and combines them into a single school Accountability Index. Schools that achieve adequate gains on the index receive financial rewards; consistent failure triggers state intervention. The cognitive measures are primarily performance-based, e.g., they include on-demand constructed-response questions and performance events as well as portfolios.

Educators in the state report that KIRIS has had strong effects on curriculum and instruction. External evaluators invited by the legislature to review KIRIS raised serious concerns about the quality of the measures and their validity for the state's purposes. This is the first example of a strong statewide accountability system built on performance measures that has been implemented, and it has interesting lessons for all educators.

Description and Purpose

The Kentucky Educational Reform Act of 1990 (KERA) represented a dramatic reform of the state's educational system, with a strong emphasis on accountability. KERA embodied a particular approach to education in that it:

- set goals for the educational system,
- created a mechanism for assessing progress toward those goals, and
- established rewards and sanctions for schools based on improvements (or declines) in performance.

There are six major goals for learners, in the areas of basic communication and math skills, applying concepts and principles to real-life situations, self-sufficiency, school attendance, school drop-out/retention rates, elimination of

barriers to learning, and the transition from high school to work or further study (see Figure A.1). Schools are the basic unit used to measure performance in Kentucky. The state expects schools to steadily improve their performance relative to these six goals.

1. Students are able to use basic communication and mathematics skills for purposes and situations they will encounter throughout their lives.
2. Students shall develop their abilities to apply core concepts and principles from mathematics, the sciences, the arts, the humanities, social studies, practical living studies, and vocational studies to what they will encounter throughout their lives.
3. Students shall develop their abilities to become self-sufficient individuals.
4. Students shall develop their abilities to become responsible members of a family, work group, or community, including demonstrating effectiveness in community service.
5. Students shall develop their abilities to think and solve problems in a variety of situations they will encounter in life.
6. Students shall develop their abilities to connect and integrate experiences and new knowledge from all subject matter fields with what they have previously learned, and build on past learning experiences to acquire new information through media sources.

Figure A.1—Kentucky's Six Learner Goals

The Kentucky Department of Education was charged with creating a system to measure and report school performance against these goals, and the Kentucky Instructional Results Information System (KIRIS) was the result. KIRIS scores are made up of two components, cognitive measures and noncognitive measures. The noncognitive measures (which account for about 16 percent of the total score for a school) include rates of attendance, retention, drop-out, and transition. The cognitive measures, which are collected in grades 4, 8, and 12, cover the core academic subjects, including mathematics, science, social studies, humanities and the arts, as well as practical living and vocational studies. Standards for performance have been set for the cognitive measures, and students' work is classified into one of four performance levels: Novice, Apprentice, Proficient, or Distinguished.

Most of the cognitive measures are performance-based,[1] including open-ended items, performance events, and portfolios. The open-ended items include both

[1] The number of multiple-choice items has been cut in half from 1991–92 to 1993–94 while the number of open-response items doubled.

short-answer and essay formats. Performance events, which last about one class period, include some group work followed by individual work leading to an individual written product. Performance events are administered on a matrix-sampled basis, with each student working on just one or two events. In addition, portfolios are collected in writing and mathematics. Each portfolio contains five to seven "best pieces" of student work that cover different content areas and different core concepts. There are no content requirements for the portfolios, but they are supposed to demonstrate breadth as well as higher-order skills in each domain.

Measures from all domains (cognitive and noncognitive) are combined into a single Accountability Index (for each school). The relative weights assigned to the content areas for the next cycle of accountability are summarized in Table A.1. A baseline Index was computed using 1991–92 performance, and an improvement threshold was established using this score. The schools' standing in 1991–92 determined its target for improvement in 1992–94. (Greater gains are expected for low-scoring schools on the baseline Index than for high-scoring schools.) Subsequent biennial averages are used as baselines for future improvement targets. Kentucky's long-term goal over a 20-year period is that all schools will score above the 100 level (which is equivalent to having all students at the Proficient or Distinguished levels).

Table A.1.

Accountability Cycle II Index Weights

Content Area	Weight
Mathematics	14%
Reading	14%
Science	14%
Social Studies	14%
Writing	14%
Arts and Humanities	7%
Practical Living/Vocational Studies	7%
Noncognitive Index	16%

Kentucky has a strong commitment to inclusion, and very few students are excluded from participation in the assessment. Special education students complete a special alternative portfolio based on their individual educational plan. Scores from these students are included in the computation of the school's Accountability Index.

Relationship to Other Programs

The Kentucky Educational Reform Act created the framework for a new educational system described that incorporated the six goals shown in Figure A.1. KIRIS is the measurement and accountability system created to support KERA. KIRIS is conceived of as one part of the "complex network intended to help schools focus their energies on dramatic improvement in student learning" (Kentucky Department of Education, 1995b). The state's goal is to create an integrated program of assessment, accountability, curriculum reform, and staff support. Because there are high stakes attached to performance, education officials expect to observe "teaching to the test," so they tried to design an assessment system based on events that were worth "teaching to."

KIRIS was built to assess school performance against the six broad learner goals as shown in Figure A.1. The Act also required the Department of Education to create a performance-based assessment program to measure success. Goals 1, 2, 5, and 6 address the application of cognitive skills, and the contractor responsible for developing KIRIS worked with educators in Kentucky to develop assessments that measured these cognitive outcomes. The learner goals themselves are too broad to serve as test specifications, so in 1991 the State Board of Education adopted a more detailed set of valued outcomes that described in greater detail the skills learners should possess in the fields of mathematics, science, art, humanities, social studies, practical living, and vocational studies.

For the next two years these outcomes were used as the basis for developing assessment tasks. However, these outcomes proved to be confusing to many important audiences, including parents, and they were replaced by a set of 57 Academic Expectations. These expectations describe what Kentucky students should know and be able to do when they graduate from high school. Subsequent KIRIS assessment development has focused on these Academic Expectations.

KIRIS was built to assess school performance in response to broad new demands placed on education. The associated outcomes or expectations were derived by panels of educators to reflect this new direction, not to articulate with existing programs. In particular, the vocational outcomes are quite general and do not necessarily match with the objectives of particular vocational programs. Only 3 of the 75 academic outcomes relate to vocational studies. These are:

- Students use strategies for choosing and preparing for a career,
- Students demonstrate skills and work habits that lead to success in future schooling and work, and

- Students demonstrate skills such as interviewing, writing resumes, and completing applications that are needed to be accepted into college or other postsecondary training, or to get a job.

KIRIS is not focused specifically on assessing learning in vocational classes. In both 1992–93 and 1993–94 only 3 performance events and 11 open-response items per grade level were used to assess practical living and vocational studies combined, and this content area counted for only 7 percent of the overall accountability index. Most students completed only one performance event and one open-response item in this domain. This does not provide enough information to be useful for evaluating vocational programs, either at the individual or program level. Over time, one might expect to see greater coordination between specific instructional activities and the statewide assessment. Furthermore, the career skills measured by KIRIS might be useful indicators of one aspect of vocational education. However, as presently conceived, KIRIS itself will not be sufficient for evaluating specific vocational programs. Rather, vocational educators may be able to learn about performance-based accountability systems from the KIRIS model.

Implementation and Administration

The state has supported the implementation of KIRIS with extensive teacher training and technical assistance. The state established eight regional service centers to train district staff as Associates, who would help their districts further professional development. Districts and schools report that the centers are a valuable resource. The Department of Education funded a program to train KERA Assessment Fellows who would be available throughout the state to help schools and districts prepare for and interpret KIRIS; over 300 educators have participated in this program. Over 100 Distinguished Educators have been trained to help schools succeed (particularly those whose scores are low). The Kentucky Educational Television network broadcast 14 professional development sessions. In addition, colleges and universities in Kentucky offered courses and contracted with individual districts to train teachers on the new assessment methods and other aspects of KERA school reform.

The contractor responsible for KIRIS has trained 700 Mathematics Portfolio Cluster Leaders to help teachers in their area understand the portfolio guidelines and implement appropriate classroom procedures. Over 1,000 teachers have participated in Guided Scoring Practice workshops for the Writing Portfolios. Teachers also have been involved in summer scoring of portfolios, which they report is beneficial for their professional development. Overall, the state has engaged in a broad and thorough effort to provide information and training to prepare teachers for the new assessment and accountability system.

Kentucky contracted with Advanced Systems in Measurement and Education (ASME) to develop and administer KIRIS. ASME worked closely with teams of Kentucky educators to formulate plans for the assessment, develop test items and open-response tests, administer the performance events, score the assessments, and set standards for student performance. ASME, in turn, contracted with Far West Laboratory for collection and analysis of the noncognitive data on attendance, retention, dropout, and transition.

It is difficult to estimate the total cost of KIRIS. The contractor receives about $6 million per year for developing the assessments, administering them, scoring the results, and reporting to schools and the state. This funding also covers some staff development activities. The Kentucky Department of Education also spends about $2 million a year on professional development of this type for teachers. In addition, some districts contract separately with ASME for additional scoring for continuous assessment, and the annual budget for rewards to schools is estimated to be about $18 million (KIER, 1995a).

In addition, the KIRIS assessment requires some amount of school time, also a limited commodity. Each student completes four periods of on-demand assessment (periods were 90 minutes long in grades 8 and 12, and 60 minutes long in grade 4). If students need additional time they are given a half-period more to complete the activities. Each student also devotes one period to a performance event, which was administered at the school by ASME staff. Writing and mathematics portfolios are collected throughout the year, but we were unable to find an estimate of the additional time spent preparing the portfolios (above and beyond the time required to do the assignments).

In addition, teachers devote some class time preparing for KIRIS; whether this is a cost or a benefit depends on the nature of the activities. KIRIS is designed to promote changes in curriculum and instruction, and, in theory, the time schools devote to preparing for KIRIS can be considered instructional time. Surveys administered by RAND suggest that teachers put a lot of time into test preparation (Koretz, personal communication). However, there is little evidence whether this was appropriate preparation (i.e., activities that promote improvement in the broad domain of skills measured by KIRIS) or whether teachers were spending time narrowly preparing students for specific KIRIS tasks or activities that might not generalize beyond the particular content of the test.

Technical Quality

In 1994 a panel of distinguished measurement specialists was appointed to investigate the technical quality of KIRIS. Their specific charge was to determine whether the Accountability Index was sufficiently robust to support how it was being used. The panel concluded that KIRIS is "significantly flawed and needs to be substantially revised" (Hambleton, et al., 1995, page 1). The panel members were particularly concerned that the public was being misinformed "about the extent to which student achievement has improved" and about the "accomplishment of individual students" (Hambleton, et al., 1995, page 5). They based this conclusion on evidence relating to six aspects of KIRIS.

All six are important considerations in the use of alternative assessment in vocational education, and each will be discussed briefly in the following paragraphs (much of this discussion is adapted directly from Hambleton, et al., 1995).

The Assessment Development Process

The greatest weakness in the development and documentation process that the panel found was that the specifications (frameworks) do not communicate clearly what students are expected to know and be able to do, and therefore they do not provide adequate signals to teachers or to test developers. Since the test emphasizes cross-cutting themes rather than traditional discipline-based knowledge, an understanding of the exact nature of expectations is important. In Kentucky the test frameworks vary in detail and specificity across subjects, and frequently they do not contain any information about variations in expected student performance across grade levels. It is important to note that the greatest weaknesses in this area were found in the first year, and the process has been improving since then.[2] The panel was also critical of the process that was used to develop assessments, recommending that the state clearly follow four steps: specify goals explicitly, construct exercises that measure progress toward these goals, evaluate the exercises by having judges examine pilot test results from students, and select and assemble test forms using acceptable items.

[2] Unfortunately, scores from the first year helped to establish each school's baseline performance level, so the initial weak test development process affected later rewards and sanctions.

The Reliability of the Accountability Index

A second problem was that the scores reported for schools did not have adequate reliability for accountability purposes: the scores reported for students were less reliable than the usual standard for such tests. The panel concluded that a substantial number of schools probably were assigned to the wrong reward category (Eligible for Reward, Successful, Improving, Decline, In Crisis), and that such errors of assignment were particularly likely for small schools. Furthermore, there was inadequate information to determine the likely level of error due to differences in task sampling from year to year, so the problems the panel was able to identify probably underestimated the true error of classification. Another problem is that student score reports do not convey information about the margin of error of reported scores, which should be included, according to accepted test standards.[3] The panel notes that reliability of both student scores and school scores (i.e., information used for assessment purposes and for accountability purposes) could be improved by using both multiple-choice and open-response tasks to obtain scores, an option that was rejected by Kentucky in its commitment to emphasize performance assessment.

The Portfolio Scoring Procedures

The panel examined separately the scores generated by the portfolio component of KIRIS, and they also reported negative findings about the reliability and validity of these scores. It is important to remember that the Kentucky portfolios served dual purposes: to provide measures of student achievement for the accountability system and to encourage changes in curriculum and instruction. On the first point, the panel found that scores were insufficiently reliable to support their use for accountability. Specifically, although raters were moderately consistent in ranking students' work, they disagreed about the percentage of portfolios reaching each of the KIRIS performance levels. More damning was the fact that ratings by students' own teachers were higher than ratings by independent judges.

There was little evidence available about the validity of scores, but the panel was particularly concerned about the lack of standardization in the way portfolio entries are produced and the amount of assistance students receive. This is a problem that undermines the validity of portfolio scores in other states, as well. Another problem of interpretation is that portfolios constructed of "best pieces" may not reflect sustainable levels of performance under normal conditions. The

[3] *Standards for Educational and Psychological Testing*. (AERA, APA, and NCME, 1985).

panel was more optimistic about the potentially beneficial effects of the portfolios on curriculum and instruction. Little information had been gathered about instructional impact at the time of the review, but, based on evidence from other portfolio assessment systems, the panel encouraged Kentucky to maintain the system on a low-stakes basis while gathering evidence about its long-term effects on classrooms.

Making Scores Comparable Across Years (Equating)

Next, the panel tackled the difficult question of the comparability of scores over time. KIRIS allocates rewards and sanctions on the basis of comparisons between performance in baseline years and in subsequent years. Therefore, it is essential that the scores be comparable from one administration to the next, although the tasks, events, and items may vary. Although much of the panel's analysis was highly technical, involving the appropriate statistical equating designs, its conclusions were clear: the equating process was insufficient. KIRIS used too many judgmental procedures without adequate standardization, particularly in the translation from raw scores to performance levels. This introduced errors into the year-to-year comparisons. Other problems that undermined the equating of scores from year to year included changes in procedures and the exclusion of multiple-choice items (with higher reliability) from the accountability index. Overall, the panel found that the equating did not support year-to-year comparisons, and it recommended a number of changes to strengthen the process.

Setting Performance Standards

Classification of students into proficiency levels is at the core of KIRIS, and the accuracy of these classifications affect the accuracy of each school's Accountability Index. Students are classified as Novice, Apprentice, Proficient, or Distinguished on each assessment, based on their scores. The assignment of scores to proficiency levels is done through judgmental processes in which panels review student responses and classify them according to descriptions of performance at the four levels. The panel found that these processes were not adequately described and appeared to lack appropriate standardization. It particularly criticized the standard-setting process, which at times assigned students to a proficiency level on the basis of as few as three test items.

The Impact of KIRIS on Student Learning

Finally, the panel looked at the evidence of educational improvement in Kentucky; in other words, has KIRIS had the desired effects on student performance? The Kentucky Department of Education trumpeted the improvement in student scores from 1991–92 to 1993–94, and the general public was led to believe that substantial progress had been made. The panel tried to determine to what extent these score changes reflect real differences in student learning. It concluded that the reported gains "substantially overstate improvements in student achievement" (Hambleton, et al., 1995, sec. 8, p. 2). Panel members base this judgment on external evidence about student performance, such as NAEP, which does not show any improvement over the same time period (although there is a limit to how many such comparisons can be made at the same grade level and for the same subject). Though the panel cannot explain these differences, they suggest that inflated gains are due to two factors: the high stakes attached to KIRIS led to inappropriate teaching to the test, and the desire to show big increases in scores led to overly poor performance during the baseline year.

Consequences and Use of Assessment Results

The Accountability Index was used for the first time in 1994 to reward and sanction schools. All schools received detailed reports of student performance and the school's overall Accountability Index. Additional money was awarded to schools that met the threshold for rewards. The reports have been used in a variety of ways that are "consistent with the intent of KIRIS" (KDE, 1995, page 222), including to monitor programs' progress over time and to target instructional program improvement efforts.

KERA and KIRIS have had broad effects on curriculum assessment, and professional development. There is clear evidence that some teachers are changing instructional practices in response to KIRIS assessments content and processes. For example, the use of writing portfolios has led to an increased emphasis on student writing. However, there is evidence that teachers are lagging in reforming many practices including some assessment-related ones. They are "struggling with the use of learning centers and theme-centered units; are failing to use recommended practices in science, social studies and the arts; are not planning their instructional program around Kentucky's Learning Goals and Academic Expectations; are having difficulty implementing a variety of continuous, authentic assessments; are neglecting to plan with special area teachers; and failing to involve parents in the primary program" (KIER, 1994, pages xvii–xviii).

Applicability to Vocational Education

Much can be learned from KIRIS that has value for vocational education. On the positive side, some of the changes that proved most difficult for Kentucky educators should be relatively easy for vocational educators who are already used to using performance as a basis for assessment. Similarly, the development of clear descriptions of desired outcomes and student proficiencies that has proved so difficult in Kentucky is very much like the task analyses that are common in vocational education and so should create fewer problems. When vocational educators try to design assessments to measure unfamiliar skills and performances (e.g., generic skills, such as teamwork or understanding of systems), they will face similar problems of definition and communication, but their experience with task delineation and performance specification should stand them in good stead.

On the negative side, strong accountability requirements seem to make most aspects of assessment more difficult. Greater resources will be needed for everything from development to training to implementation if such an assessment is used to structure an accountability system.

None of the assessment elements of KIRIS is new; other testing programs use portfolios, performance events, and open-ended responses, and other states produce school "report cards" with indicators of both cognitive and noncognitive outcomes. What is unique about KIRIS is the use of these measures in a strong accountability context. The presence of high stakes exacerbates the political problems, raises the necessary technical standards, and heightens the anxiety level of educators, all of which would make it difficult to implement KIRIS-like assessments in similar contexts. The use of a single summary index of performance without the high stakes is one that might be beneficial for some purposes, however.

Of particular concern is the need for high-quality measurement, a goal that still eludes KIRIS after four years (according to the technical experts). Such quality standards increase the time and resources needed for all aspects of the assessment, including developing student outcome goals, producing assessment specifications, developing tasks, scoring student responses, setting standards, equating forms, and reporting. Such technical issues will have to be confronted by vocational educators if they want to use performance assessment for certifying competency, awarding certificates of mastery, or other important uses. In fact, the technical demands will be greater if the assessments are going to be used to make decisions about individuals. The KIRIS experience suggests that such an approach will require advanced technical expertise as well as considerable time and resources.

B. Laborers–AGC Environmental Training Assessment

In 1969, the Laborers International Union of North America and the Associated General Contractors of America (AGC) established a cooperative trust fund for the common purpose of improving the skills of construction laborers. The union sought to increase the demand for its workers, the contractors wanted more productive craft workers, and both parties had a vested interest in creating safer workplaces. For the last 26 years, the Laborers–AGC Education and Training Fund has been meeting these goals by developing and supporting occupationally focused courses for 66 affiliated local training schools in the U.S. and Canada. These schools are responsible for training the 350,000 union members (half of the membership) who work in construction or environmental clean-up. More than half of the schools (40 out of 66) offer environmental courses in addition to construction offerings. Contractors pay money into local trust funds. This money is added to each worker's total benefits package; these funds pay for running the affiliated schools and help defray Laborers–AGC's costs for curriculum and assessment development and technical support. As an example, the Northern California training school has an estimated operating budget of $700,000 a year, and of the $0.21 per worker-hour that local contractors contribute to the training fund, $0.02 is sent to Laborers–AGC for support.

The first 15 years of the Fund's efforts concentrated on general construction safety programs and courses on specialized areas of the industry. Though they were developed with union funds and for union members, some of Laborers–AGC's films and course materials were used by U.S. and Canadian government agencies for worksite safety and awareness programs. In the mid-1980s, the Fund shifted some of its efforts away from construction. Union officials noticed a significant lack of organized workforce development in the burgeoning environmental remediation industry. Labor market projections at that time exposed a potentially severe shortage of skilled environmental workers. In 1987, Laborers–AGC received a grant from the National Institute for Environmental Health Sciences (NIEHS) to develop a program for hazardous waste clean-up workers. Favorable program evaluations led to further grant awards from the EPA, U.S. Department of Energy (DOE), U.S. Department of Defense (DOD), and the National Institute for Occupational Safety and Health (NIOSH).

NIEHS and NIOSH distribute grant funds and monitor administrative requirements, but they rely on DOE and DOD for technical standards and evaluation. These agencies are each responsible for particular environmental areas and must regulate all training programs that certify workers for these fields. All environmental workers must be certified to work and all training programs must be formally approved to operate, because these workers handle substances that pose potentially serious risks to public health and safety. The Fund's program specialists, who develop the course curricula and train course trainers, must do so in compliance with the mandates of the federal agencies that oversee each work specialization.

In addition to meeting the requirements of the federal agencies, programs often must meet additional state agency requirements. The differences not needed among the various state and federal standards make it very difficult for Laborers–AGC to achieve programmatic consistency among its environmental courses. Each course (shown in Table B.1) is independent and leads to a specialized certificate, but the Fund maintains a single approach for all these differing courses. One standardized element in these courses has been the assessment system used in the environmental courses.

Table B.1

Environmental Courses

Course	Hours
Hazardous Waste Operations	45
Hazardous Waste Worker	80
Asbestos Abatement	40
Lead (paint) Abatement	40
Radiological Worker	32
Underground Storage Tank Removal	32*
Confined Space Entry	32*

* New courses with hands-on activities and written tests, but no formal performance assessments yet.

Description and Purpose

The Laborers–AGC Education and Training Fund's environmental training assessment is a flexible system that uses performance-based tests and criterion-referenced multiple-choice tests to measure the competencies and knowledge of environmental trainees. The assessments are designed to certify each individual's competence, as well as to monitor and report program completion

information to the appropriate federal and state agencies. The programs receive federal funding, so the latter use of results is done both to comply with governmental certification requirements and to maintain quality standards and accountability for the ongoing grants or contracts.

The Fund developed each assessment tool by employing an assessment expert to work with each course's program specialists and industry experts. In some instances, staff from regulatory agencies were consulted on specific issues. The cost for developing the written and performance assessments was $10–12,000 for each course, all of it covered by grant funds. Laborers–AGC staff attributes the relatively low development costs to high in-kind contributions from training school instructors and assessment specialists. Also, as previously discussed, the Fund is just now working to rigorously evaluate the assessments for validity and reliability.

Courses range from 32 hours to 80 hours in length. Students spend roughly half their time in the classroom and half in hands-on field activities; usually they must pass all of the performance tests in order to continue in the course and to be eligible to take the multiple-choice exam given on the final day. When they successfully complete both, trainees gain Laborers-AGC–sponsored certification and can work for up to one year in the particular job.

Courses have from 1 to 6 performance tests, depending on course curriculum and length; each test may assess up to 35 tasks. The performance tests last from 5 to 20 minutes; the simplest requires a trainee to explain his or her actions while testing certain equipment. On the other end of the spectrum, the trainee may perform a complex series of actions in a simulated work procedure. In some courses, the performance tests are distinct events that occur separately from the regular training, while in others these assessments are used as a training tool and then later as a measurement tool. In the latter case, trainees pair off during the training event, one evaluating the other's performance using a check-off sheet. The instructor monitors the evaluations, with little interference, and uses the same check-off sheet to test them later. This shared evaluation helps trainees build a stronger sense of responsibility for their own knowledge and performance as well as for that of their coworkers, on whom they will rely so heavily at the work site.

In the Hazardous Waste Worker Course, one procedure that is both a training activity and a tested event is decontamination after simulated field work in a Level–A protective suit. Often called a "moon suit," the Level–A fully encapsulates workers and their protective equipment (boots, hard hat, respirator, and air tanks). A trainee enters the three-stage "decon" area wearing the suit and

proceeds through a battery of prescribed steps for washing and disrobing. He or she must first make sure to properly spray and scrub his suit with disinfectant before moving to the disrobing stage, when the trainee removes each layer of protective clothing and equipment. Trainees take approximately 15 minutes to perform all 19 steps involved in the decon, each of which must be performed properly and in sequence to pass. There is no limit to the number of re-tests if trainees fail this procedure, since it must be passed to pass the course.

The performance test criteria vary from program to program, according to the degree of oversight by the regulatory agency. For example, the Radiation Worker Course must meet carefully specified Department of Energy regulations. The performance exam for this course, then, utilizes importance-weighted point deductions for incorrect performance on tasks. For example, if a trainee fails to remove protective gloves in the proper sequence, two points will be deducted, but if he or she improperly responds to an "unusual radiological event," five points are deducted. The underlying factor that determines each task's point weighting is the potential for health and safety risks to the individual, coworkers, or the public if the trainee makes a mistake. Of the performance test's 23 tasks, three carry possible deductions of 21 points each—of themselves, enough to fail the test—because these are crucial tasks that workers must *never* perform improperly. If an individual's point total drops below 80 on the assessment, he or she cannot continue the course. For every task, though, instructors have a box to check if the student recognizes his or her mistake just after making it, notifies the instructor, and rectifies it immediately. The point deductions decrease when trainees correct themselves this way, and, in the case of the three crucial tasks, the decrease is dramatic—from 21 points to 7 or even 3 points (see Table B.2). The Fund developed these weightings with input from and approval by the Department of Energy.

By contrast, the Hazardous Waste Worker Course does not have such a rigorous level of performance evaluation. The Occupational Safety and Health Administration (OSHA) requires workers in hazardous waste removal to be certified by an approved training provider, but it does not produce regulations that specify what the training must include. Laborers–AGC is left to define the requirements for certification, including designing the performance exam in this case. Each item on the performance exam is simply marked as correct or incorrect. Though it would be possible to generate importance weightings for each task on this course's exams, Laborers–AGC would have to shoulder the costs of researching them and then justify the weightings to a federal agency that does not even require a performance test.

Table B.2

Sample Items from the Core Radiological Worker Training Performance Test

TASK	RATING			
	Performs Correctly		Notifies Instructor	
	Yes	No	Yes	No
4. Recorded correct information for task on RWP sign-in sheet prior to entry.	____	____	____	____
		-3	-2	-3
5. Entered only areas identified for tasks on RWP	____	____	____	____
		-21	-7	-21
6. Maximized distance, minimized time, and utilized shielding.	____	____	____	____
		-5	-3	-5

The written exams are given at the end of each course and consist of either 50 or 100 questions that are drawn randomly from a large test bank. Laborers–AGC creates the questions for each test and submits them to the appropriate federal agency for approval. The Radiation Worker course is an exception, though, in that Laborers–AGC staff must randomly select questions from those developed by the Department of Energy. Once tests are formulated, they are disseminated to the training facilities, where local instructors administer them according to program guidelines.

Results of the written exams are tallied for each individual and later aggregated for whole classes, training schools, and the entire training system. Local training schools need student and class results in order to process worker certificates and to comply with state or local regulations for training providers. Laborers–AGC collects all data to monitor both of these processes and to keep track of program performance trends.

The assessment tasks are tied closely to the instructional objectives. These objectives were developed by the Fund's content/industry experts and agency staff to mirror the skills needed in the occupation. The hands-on activities contextualize the classroom information in events that will be found at most, if

not all, environmental remediation work sites. These field activities use mock hazard sites and actual tools and equipment, to ensure that once trainees pass a performance event, they can work safely and effectively on real work sites.

The knowledge and skills measured by the assessments are highly specific, to the occupation and the specialization area. At present, Laborers–AGC is considering ways to combine courses that are regulated by different agencies to create more comprehensive environmental worker courses. Though such courses might have great potential for workers and employers, the Fund finds it very difficult to satisfy all the current state and federal regulations simultaneously for each individual work area. However, it combines two training courses regulated by OSHA and DOE (Hazardous Waste Worker and Radiological Worker, respectively) into a single 120-hour course for workers at nuclear power plants. In January, 1996, program specialists were preparing for a trial run of the course at the Hanford nuclear facility in Washington state. Any significant alterations to these assessments or the curriculum will occur only after a review of course results and input from the agencies. Laborers–AGC administrators are considering how to combine the EPA–regulated Asbestos and Lead Abatement courses into such a "cluster" course, but considerable work with EPA staff will probably be necessary to do so.

Using the Fund's curricula is optional for local training schools, but the Laborers–AGC programs have been independently approved by the regulatory agencies, so it is a definite advantage for local schools to use them. Documenting that they use the Fund's federally approved program helps schools satisfy most, if not all, of their state's requirements for providers of training in these specializations.

Relationship to Other Programs

Many union laborers consider the environmental training courses useful for career advancement. In large numbers, construction laborers seek environmental certifications after working for several years in the construction field. Union members generally agree that environmental courses are more technical and have more formal testing procedures than most construction courses, thus requiring greater cognitive abilities. The nature of these courses led to developing a preparatory course for union members who want to bolster their basic reading, math, and science skills before enrolling in an environmental course. The preparatory course lasts 40 hours, uses some texts and materials from the certification courses, and is usually offered just before many environmental classes start, so that these trainees can quickly apply their sharpened learning skills.

Although the Fund's environmental programs are now equal in number and importance to its construction programs, there is little contact between them. They operate under separate departments and, in general, have separate sources of financial support. The construction programs use mainly local training fund contributions, while the environmental programs are supported by federal grants. The environmental courses have money to support activities such as hiring consultants to develop assessments and evaluate programs. The construction programs are less able to do this, but recently the Fund's administrators have undertaken an initiative to research and create new performance tests for them. However, until this project is completed, most of the construction programs will continue to use informal instructor observations as the sole means of skill assessment. Environmental and construction programs differ in their level of technicality and certification requirements, and there is less overall consistency among the construction courses. Within the environmental department, though, courses are closely related. Program specialists often cross-train so that they can collaborate on curriculum development and train-the-trainer events. This collaboration helps to increase consistency of training in specialized areas that may be technically dissimilar.

There are many other environmental programs that prepare workers for this field, but coordination or cooperation between them is rare and limited. Other unions such as the Carpenters, Teamsters, and Operating Engineers offer certification programs in the same specializations, as do many private organizations and postsecondary institutions such as the University of California at Los Angeles. Competition for students is strong among all these organizations. First, private training schools and postsecondary institutions compete for students: the first to make profits, the second to fill enrollment targets. Second, the unions compete among themselves to place more of their workers in the courses and then in jobs than the others, thereby gaining more of the market. This latter tension is difficult to resolve because under the NIEHS grant, Laborers–AGC is the primary grant recipient and the Teamsters Union is a subgrantee. Each is developing independent course curricula, but the Fund has additional duties. Laborers–AGC is responsible for all administrative, budgetary, and reporting concerns. Interaction between the two is mainly confined to high-level administrative matters, and staff members do not confer frequently on curricular matters.

Some of Laborers–AGC's programs differ from courses offered by other providers in that they require more, sometimes double, the course hours than the regulations mandate. Industry consultants recommended to the Fund added time for extensive field exercises and assessment in addition to classroom

instruction. Not all of these hands-on activities are required by federal agencies, but many have real safety and productivity implications. While Laborers–AGC's Hazardous Waste Worker Course is recognized in the industry and agencies for its quality and comprehensiveness, the 80-hour course time may be a disadvantage. All the other unions and private institutions that offer this worker certification do so in a 40-hour format, appealing to those paying for the training since it costs less and responds more quickly to employers' requests for qualified workers. Contractors with large clean-up projects do not compare course quality or assessment procedures when trying to meet workforce needs and project deadlines. They simply need certified workers and may call upon another union if it can supply them faster and at lower cost. Due to this pressure, Laborers–AGC is considering designing another version of their field/classroom training in a 40-hour format. Currently, they do not know how this change would affect the assessments.

In addition to the environmental course assessments, trainees in the Hazardous Waste Worker course must successfully pass a physical exam in order to participate. In the physical exams, a registered nurse tests each person's pulmonary capacity, heart rate, and blood pressure. The physical exam is given outset of the course, to provide assurance that each trainee has the physical capacity to perform strenuous training activities (and, later, work) in enclosed suits while wearing respiratory protection. Along with the signed approval of their physicians, this assessment's results are recorded as part of each person's eligibility for training and subsequent certification. This reduces the Fund's and the training schools' legal liability for any incidents that may occur as well as screening these individuals before they begin the course.

Implementation and Administration

The traditional model of assessments, multiple-choice final exams, has been used in one or more courses continuously since 1987. As the courses have been developed and come on-line, the assessments have been adapted slightly in order to reflect the standards and certification requirements of each course. Responsibilities for the environmental assessments are divided among staff at Laborers–AGC and staff at the 40 training schools that offer at least one environmental course. Program specialists at Laborers–AGC develop and update the tests as well as monitor the quality and consistency of their use at local sites. The Fund's Director of Environmental Programs is responsible for overseeing all assessment and other curricular activities. Training-school staff administer the assessments, score them, and report the results to the Fund and state regulatory agencies.

The environmental assessments are updated by the Fund once a year, or more frequently if significant changes occur in the industry or its regulations. When considering changes to the assessment, the Fund relies on the expertise of its specialists and other industry or regulatory experts, as well as input from course instructors. At its annual Instructor Development Program (IDP), the Fund holds educational seminars on professional, technical, and life skills topics for the more than 200 instructors. Also at the IDP, the Fund holds curriculum update sessions for each environmental course, where instructors can discuss issues directly with specialists, in order to maintain course integrity at the local level.

Technical Quality

Laborers–AGC has not extensively evaluated the assessment tools used in its environmental programs. In the late 1980s and early 1990s, much of the Fund's efforts concentrated on developing and disseminating courses to meet the training demands of employers and the union. One course after another was developed and brought on-line throughout North America. Due to staffing and time constraints, extensive reliability and validity checks were not performed during this period. In the last few years, though, the Fund has started efforts to evaluate and strengthen the technical quality of its environmental assessments.

Laborers–AGC staff, together with technical and assessment experts, began first by reviewing the oldest assessments—those from the Hazardous Waste Worker Course. Though Fund staff originally developed the tests for this course with the guidance of similar experts, experts did not remain involved throughout the development process, which may have contributed to test weaknesses. After lengthy evaluations of the written test, reviewers found items that did not comply with best-practice guidelines for multiple-choice criterion-referenced exams. The Fund set out, with assessment specialists, to remedy the problem items by creating a bank of draft test questions that met the guidelines. These questions were then screened by subject specialists for content validity and by assessment specialists for construct validity. The resulting questions were used in pilot course trials at several local training schools. Work is currently underway to synthesize the collected feedback from course instructors, students, and program specialists so that final changes can be made to these test items. Once reviewed and corrected, these items will be incorporated into the current test bank, and the same process will be applied to written exams for the other environmental courses.

All the performance exams will eventually undergo comprehensive evaluations, but the Fund has not yet determined the process for this. Only the Hazardous Waste Worker Course's performance tests have undergone a preliminary evaluation. Content and assessment specialists found that test items are strongly correlated with the work performed on actual work sites, though it is clear to the Fund that continual changes in technology, materials, equipment, and practices makes content validity an ongoing concern. The items tested in the performance exam were found to closely reflect the course content (as reflected in curricular materials), but in some instances they did not closely follow what was actually being taught. For example, certain items in the performance test, as in the written exam, are meant to measure students' ability to integrate situational facts and circumstances to arrive at a proper solution or action. In some course-monitoring visits, reviewers found instructors were not properly teaching the skills needed to do this. The situational facts were covered, but instructors often did not lead students through the synthesis steps of linking background information and circumstances with possible actions and their likely impacts.

Laborers–AGC considers this flaw both programmatic and instructional, and is working to strengthen both the assessment skills of its program specialists (who develop the curricula and train instructors) and the instructional skills of its trainers. Fund staff work with each other and assessment consultants to understand how to develop curricular activities for these skills, and they work with small groups of instructors in yearly instructor refresher sessions to ensure that the skills are taught properly. The work that Fund specialists and other staff plus consultants do on curricular and testing updates can be further refined and coordinated at the Instructor Development Program, where the entire cadre of environmental instructors gathers yearly.

The American Council on Education has also evaluated these environmental courses through its Program on Noncollegiate-Sponsored Instruction (PONSI), though to a lesser degree. Instructional and subject experts representing PONSI compared each environmental course's content, learning activities, and assessments with current college offerings. Each course was given a recommended number of semester-hour college credits. PONSI re-evaluates each course every five years, or sooner if course components are changed. This continuing evaluation is another source of maintaining high quality in the environmental assessment system.

Employers' reaction to the quality of certified employees is very positive, especially notable because employers are mindful of the potential health and safety ramifications of improperly trained workers. Because the Fund is a shared venture between labor and management, employers have immediate input

channels if their needs are not being met. The construction and remediation contractors are not the only employers who rely on the assessments to accurately measure skills, though. The DOE, for example, has contracted with the Fund to train workers at its headquarters and at several nuclear facilities. DOE experts take great interest in this training because the Department requires managers and some engineers at nuclear facilities to be certified, as well as the facility technicians. Reaction to the skills assessment from all levels of the Department and participating employees has been positive, just as it has with industry contractors and union members. When informal pre- and post-training comparisons of workers were conducted, in interviews with union members and DOE staff, they showed improved knowledge, awareness, and overall performance.

Since the first generation of tests, Laborers–AGC has monitored the exam for any form of gender or racial bias and has made changes when necessary. For the most part, questionable test items are discovered either in monitoring visits or at curriculum update sessions at the IDP each year. Though the union membership is roughly half female or minority, assessment results are not aggregated by gender or race-ethnicity to allow such comparisons.

Consequences and Use of Assessment Results

An obvious consequence of the assessment is a certificate specifying skill achievement, and acceptance into a specialized field for those who earn it. As a result of certification, some trainees have greater confidence in their own skills and knowledge, and they gain greater awareness of the potential effects of their actions on the job. The performance assessments in particular give them the capability to monitor their own work performance and the safety conditions affecting them and their coworkers. At their yearly certification refreshers, many trainees have commented that they mentally "test" their performance while working and, as a result, feel safer and more sure of their decisions in the field.

On the negative side, some potential students who doubt their classroom skills (e.g. technical reading, listening for comprehension, etc.) have considerable reservations about enrolling in these courses. Word-of-mouth accounts of the written tests' difficulty in particular cause many to fear that, even if they successfully participate in all classroom and hands-on activities, they may fail the final exam. Although the failure rate (on the final exam) for the environmental courses is only about 10 percent, more than half of the students enter courses with a substantial fear of failure, which contributes to a fairly high drop-out rate before the exam.

An additional result of the assessments is that some instructors see them as yet another of the Fund's curricular mandates. Because local schools are independent, some are reluctant to comply with strict rules or to use required curricular components. Laborers–AGC staff suspect that some assessment rules are not followed from time to time (such as orally translating test questions into other languages), but such deviations are likely isolated and rare. The Fund's program specialists visit each school every 12 to 18 months to monitor particular courses for compliance. If a school blatantly disregards program rules, Laborers–AGC can take sanctions including withdrawing its sponsorship, and a school would then have to develop and accredit its own course. This would entail researching technical and pedagogical issues, developing the curriculum and materials, and purchasing new equipment and supplies, in addition to gaining state and federal agencies' approval for the program. This process would be extremely time-consuming and costly, so schools have a strong incentive to comply.

The courses cover a great deal of material and instructors must essentially teach to the test. This is seen by the Fund as both a beneficial and necessary measure, because test-focused instruction gives students an acute sense that all items in the comprehensive course are applicable and important. Laborers–AGC staff sees test-focused instruction as contributing to both the strong correlation between assessments and work performance and to consistency in instructional and assessment practices across schools.

Applicability to Vocational Education

The Laborers–AGC assessment model is one that may be applied very easily in vocational education settings. That is, a system of performance-based and written tests is not unusual in vocational settings, but two characteristics of the Fund's model would be difficult for many vocational programs to match: The high level of industry support, and the high level of funding. A key element to the Fund's assessment that would not exist in many vocational settings is the strong systemwide partnership between employers and workers, and their input to the tests (indeed, gaining substantial input from *either* employers or workers is usually an obstacle). With such regular industry input, Laborers-AGC assessments can test for federally mandated skills as well as those required by employers. The broad industry base that provides this input allows certified workers to gain skill portability and enables training centers to meet many of the common demands local employers voice. Of course, vocational educators may be primarily concerned with the demands of employers in their particular area or state, but as national skill standards are developed, the prospect of a broader, industry-validated assessment may become desirable, and even necessary, in many programs.

In light of this, an assessment system that can adapt to the various standards and regulations governing occupations is one many educators could benefit from. In vocational programs without a consistent assessment approach among courses, this system could serve as a model for linking common elements and emphasizing them in the courses, while allowing for variation between subjects. Laborers–AGC maintains the work-simulated performance tests and multiple-choice written exam in all courses, even though the nature of course content and applicable federal regulations may vary considerably.

It is also important to recognize that costs may be a barrier for vocational educators who seek an assessment with the depth and breadth of hands-on activities in the Laborers–AGC model. Space and equipment requirements for the Fund's model are quite high. It is very costly to obtain and prepare areas for practice and performance so that students can closely simulate actual environmental remediation work. The Laborers–AGC assessments rely on intricate and varied field activities to measure how a student will perform on the job. It would not be plausible to conduct the Fund's assessments in a small yard or shop bay that must be shared with other classes. Each simulation area is generally dedicated to a narrow range of tasks. For example, the Hazardous Waste Course requires trainees to perform activities in an outdoor field simulating a hazardous substance dump, wherein trainees maneuver to locate and uncover barrels buried in dirt or sunken in small pools of water. Even in the Asbestos Abatement course, where simulation areas are indoors, the curriculum calls for a dedicated room or properly enclosed structure that allows the simulated asbestos particles to be removed and hauled away, just as it would in a true remediation area.

Equally costly as creating the simulated work sites for practice and performance assessments is the use of actual equipment. The courses require enough equipment for all students to use or wear items simultaneously. In general, this equipment is very costly: respirators, air tanks, and specialized air filtration vacuums. Of course not all vocational courses prepare students for occupations that use expensive tools and equipment, but each prepares them for jobs where the equipment and settings are unique. Without using that equipment in a wider variety of situational applications, as Laborers-AGC does, vocational teachers may not create or maintain such close ties between course content, assessments, and projected job performance.

Once the performance area is established and equipment purchased, though, the cost of administering these assessments is fairly low. Instructors' time is all that is required (plus a small amount for materials), but for safety and pedagogical reasons, more than one instructor must be present if more than five trainees are

in certain kinds of protective suits or using certain equipment at one time. The same instructors who administer the tests score them using criteria or answer sheets provided by Laborers–AGC, so the cost of their labor is the main expense for scoring.

There has been considerable outside interest in Laborers–AGC's environmental programs, though mainly from educators outside the country. The Fund is currently working with industry training organizations in countries such as Mexico and Russia on implementing some of the Fund's curricula. Though many environmental problems are common among countries, the new programs and assessments will have to adapt to different government or industry regulations where they exist. Because the assessments developed and used in the U.S. and Canada were built to accommodate such differences, the Fund feels this dissemination effort will progress smoothly. Though generally minor, the adaptations necessary to accommodate for differences between U.S. and Canadian regulations will prove valuable as the Fund develops these foreign programs.

C. Oklahoma's State Competency-based Testing System

Description and Purpose

The Oklahoma competency-based testing system encompasses a range of criterion-referenced, multiple-choice and performance-based assessments that test the competency attainment of students in both comprehensive high school vocational programs and vocational technical centers (testing includes secondary and postsecondary students). The Oklahoma Department of Vocational–Technical Education (Oklahoma Vo–Tech) developed and oversees these tests; it is a separate agency from the state's Department of Education. The testing system is used to achieve three objectives:

- Program improvement and accountability at the state level (providing data for the occupational competency attainment measure in the state's Perkins performance measures and standards);
- Improving instruction and student learning through competency-based curriculum and assessment; and
- Certifying that students have attained competencies for employment purposes.

Criterion-referenced written multiple-choice tests have been developed for 190 occupational titles that are categorized into 26 program areas. A new written test is administered every year for each of the 190 job titles. Advisory groups have been established for each of the 26 areas, to create duty/task lists and to rank tasks by importance (based on how frequently they come up on the job). Questions are written using these lists and then entered into a secure test bank. State staff randomly selects test items from one of the duty areas to develop the annual tests, which require a minimum score of 70 percent for passing. Oklahoma Vo–Tech is no longer creating assessments in areas where liscensure procedures exist, such as aircraft maintenance and cosmetology. However, the State Department of Health is negotiating with Oklahoma Vo–Tech to develop and administer their liscensure exams (see discussion below).

Students are required to pass two performance assessments, attaining 100 percent of the competencies tested, before they are allowed to take the written test. The advisory groups developed performance assessments for use across the

state. Though instructors must use the statewide written test, they are free to select tasks for the performance assessment from the state-developed ones or to use their own performance assessments from their curriculum. Instructors are not required to report passing rates for performance assessments or provide evidence that students passed the tests to the state, only to keep documentation at the school site for state review and audit purposes.

Performance assessments may be administered throughout the year or at a single point in time. Written assessments are administered once during the school year but not on a set date. After taking the written test on the selected day, students may retake it as many times as the instructor allows. There are multiple test forms available from the test bank to allow pre- and post-testing, and testing of the same students in successive years.

Results on written assessments are reported to the individual students and instructors. Testing liaisons/assessment coordinators at the school site receive a report that describes the performance of the program as a whole in each duty area. Scores are not used to compare different programs (they are aggregated only to compare programs to a standard). In fact, neither superintendents, principals, the state director, nor the assistant director can access individual or program aggregate scores, because the department fears they could misinterpret the data. Only the state-level program manager and her staff can access the data.

Knowledge and skills assessed are occupation-specific. State staff is currently working with advisory boards to develop assessments in occupational clusters in eight areas (hospitality, health, manufacturing, finance, agriculture, marketing, transportation, and construction). Core cluster skills are being identified for each cluster. However, an occupationally-specific test will still be administered for all 190 occupational titles. It is unclear how cluster tests and occupationally-specific tests will fit together. The first duty area for each cluster is work-readiness, where basic academic skills such as estimating are embedded in occupational test questions.

Using the task list created during test development, curriculum guides have been developed for each occupational title. Most schools use the curriculum guides, but it is optional. Curriculum guides include a post-test, which in practice is now used as both pretest and posttest in order to measure competency gains. These are reported by the testing liaison for the gain measure included in 1990 Perkins Act–mandated performance measures and standards. Scores from the written assessments are used for the attainment measure.

Implementation and Administration

The assessment system has been fully operational for the last 10 years. Advisory committees continue to meet annually to revise task lists and tests. Before new tests are administered every year, each advisory committee reviews the items on their test. The committee for each occupational title includes representatives from labor, higher education, secondary faculty, and industry.

Three test specialists (state-level vo–tech staff) are in charge of the 26 program areas. Each of these three specialists thus coordinates the advisory committee's work, test development, administration, and scoring for about 9 program areas, or about 63 occupational titles. It is very difficult for staff to keep the task lists up to current standards given this heavy workload. Task lists are thoroughly reviewed every two years.

State staff relies on testing liaisons at each school site to administer the written tests. Testing liaisons must be trained in the areas of objectivity, test security, and administration. Testing liaisons were trained in all five regions three years ago, and are now updated every August at the annual vocational conference. In addition, staff work with educators who work with student teachers, who will use the curriculum guides and tests when they become teachers. The liaisons were also given inservice training on performance assessment, but the state has no intention of centralizing that procedure. Instead, it hopes to move toward a passport system which will include portfolios to portfolios document the instructional process (and perhaps other elements).

Scantron forms used for the written competency tests are mailed in mid-April to May to each test site. Individual and group results are reported at the end of May, as well as the mean percentage correct statewide, and for each program, for each task. These scores are used by individual teachers for measuring classroom performance, and by the test liaison to report competency attainment for program areas for the statewide measures and standards. Competency gain is reported from scores on the pre- and post-test included in the curriculum guides. Unlike the written tests, the hands-on component is not secured in any test bank, nor is it administered consistently.

Relationship to Other Programs

Oklahoma has had a relationship with VTECS for the last eight years. The state agency provides VTECS with tests and a pricing structure, but it does not give VTECS access to the test bank. (The state wants the test bank to remain secure and VTECS has never been able to figure out how to keep it secure once their

members gain access to it.) Arkansas, another active VTECS member state, has an unsecured test bank that is available to VTECS members. VTECS has a direct software management system (ABACUS) that allows the instructor (or state-level staff) to enter curriculum content into the system. Oklahoma staff are interested in having such a system because they think it would make the pre- and posttest part of the curriculum guides cheaper and easier. Oklahoma needs to figure out how to maintain test security yet connect the curriculum with industry (rather than let the teacher alone decide what to teach).

The state's Vocational–Technical Education Department's assessment system also has links to the assessments administered statewide by the Health Certification and the Associated General Contractors' programs (see next section).

Technical Quality

Based on some pilot testing, the state staff and committees believe that scores on the multiple-choice test are closely linked to job performance. Because they believe there is no real need for a statewide evaluation of the performance assessment system (and because of the high cost of conducting one), they have opted not to pursue such an evaluation. The Oklahoma assessment system relies here on military research that concluded that cognitive knowledge (tested in multiple-choice format) is the best indicator of performance (knowledge transfer). The hands-on performance component is still available for use statewide, but it is not required. No formal study has been conducted to investigate correlations between the performance components and written tests. The state office does not collect data on the failure rates on performance tests conducted at local sites.

State staff does conduct item analysis on each multiple-choice test question to look for questions that are too difficult or that indicate gender or racial/ethnic bias. If they find test items that students are consistently not getting right, they throw them out. Staff members also look at the number of tests administered to each student and the number (and percentage) of students re-testing.

The state staff feels confident in their tests' content validity, given the employer input to the assessment system. The committees meet regularly to update task lists and test questions in order to ensure that the system continues to be useful in opening up employment opportunities for students.

Consequences and Use of Assessment Results

The assessment system has always been implemented on a partly voluntary basis. Local sites are directed to use the system by Oklahoma's state staff, which has more centralized control than most states (for fiscal and historical reasons), but the state cannot force schools to use the system. Charles Hopkins, the Assistant State Director of the Vo-Tech Department, wishes the federal legislation had more "teeth." There are no real consequences for not using the state-prescribed system, and the 29 districts (59 campuses) can choose which measures and standards to comply with. However, under the system of performance measures and standards required by Perkins, schools' use of the competency-based assessment system has dramatically increased. Local programs receiving federal funds are required to report their performance using these written assessment results.

However, because of this connection with Perkins, more often than not instructors view the tests as contributing to the state agency's accountability system, not as a system developed for certifying students or for program improvement. At this point, many instructors use the tests mainly to comply with the state system. Some instructors have difficulty seeing the connection between what they teach in class and what is tested on the competency assessments. One reason for this lack of connection is that the curriculum guide (which teaches to the test) is used inconsistently across schools and occupational areas. Many instructors, schools, and districts use their own texts and other curriculum components instead.

The state staff is hopeful that this compliance mentality will shift when teachers begin implementing the passport system, in which students will earn a certificate to show potential employers. These assessments' multiple-choice, performance-based, and other components will be combined in new ways so that students can earn a passport in one of eight occupational cluster areas (listed above). State staff hopes that students and instructors will see the value of earning a passport and, therefore, begin to see the competency-based tests as part of a certification process, not just meeting accountability requirements. Participation in the passport system will also be voluntary.

Applicability to Vocational Education

Oklahoma has a strong tradition of vocational education, including centralized state control and substantial state funding, which contributes to the successful operation of the assessment system. Without a strong funding base and

centralized system of vocational education sites, other states may be hard-pressed to follow Oklahoma's example. Oklahoma has both vocational-technical centers (59 campuses) and comprehensive high schools in 29 districts. Each is governed and funded by a separate structure. Through the state board of vocational-technical education, the state budget funds the entire testing program, which includes a state agency staff of about six. Test liaisons at each center are essential to the system. Most of the funding for the liaisons and for other costs of operating local test centers comes from local government sources; however, about 20 percent of the testing liaison's job is dedicated to the state's assessment system.

Oklahoma also points to the difficulty of implementing a state-driven, centralized assessment system that is perceived as relevant to the local level. Many instructors, especially those teaching courses on a narrow range of topics, have difficulty figuring out how what they teach fits in with the tests' content, because the state tests may have a much broader scope. For example, the Advanced Electronics instructor's teaching focuses on microcomputer skills and knowledge, but the tests his students take is the General Technician Test. On the other hand, a Business/Technology instructor in a Systems Management program teaches computer skills, but not for a particular job like "receptionist/word processor"; there the instruction may be somewhat broader than the area covered by the test.

Some vocational-technical centers do not use the state's curriculum. Many instructors like to develop their own curriculum and their own pre/post tests. In one center, 3 of the 24 programs use the state curriculum. Copies of the curriculum are available for reference and use. Instructors determine when students in vo–tech centers who are self-paced have completed the program (even if they do not pass the state tests, they can be called a completer).

Oklahoma Health Certification

Description and Purpose

The Health Certification Project is administered jointly by the Oklahoma Department of Vocational–Technical Education and the Oklahoma Department of Health. Certification is currently administered in three areas—long-term-care nurse aide, home health care aide, and medication aide, with assessments under development in four other areas: adult day-care program aide, residential care aide, developmentally disabled aide, and non-technical medical care provider.

Students complete a training program that is approved by the Oklahoma Department of Health and then take a two-part test:

- A clinical skills test, where candidates perform tasks related to client care.
- A written test of 70–90 multiple-choice questions.

About 5,000 students complete the assessments each year in the three areas currently in operation. An RN or LVN must approve the clinical performance part of the test, which covers three selected objectives that change from year to year (these are selected from a comprehensive list of objectives). The test liaison trains the RN or LVN to be test judges, using a guide developed by the state. Only 43 test sites can administer the written portion of the test, but any location (including a hospital) can be approved to assess clinical skills. Any person can work up to 120 days without certification in the three areas. With certification, long-term care aides are typically paid $5.25 per hour and home health aides $8.00.

Implementation and Administration

Students must pass the clinical skills tests before taking the written test. Starting in July 1995, students are required to complete 75 hours of classroom training and 12 hours of clinical training before taking either test. Students study the subjects identified in the Health Certification Project Duty/Task List developed by the Oklahoma Vo–Tech Department. Test are developed using the same method as the Oklahoma competency-based assessment system uses (described above).

There are 43 different test sites located throughout Oklahoma. These tests must meet federal and state liscensure requirements for the relevant occupation. It costs each student $30 to take the clinical skills test (the home health aide one is a little more expensive because it requires 13 competencies). The fees collected go to the area vocational-technical schools. It costs $30 to take the written test (area vo–tech keeps $5, the state vo–tech department gets $25, and the state sends the health department $3). Oklahoma Vo–Tech is breaking even in administering the health tests.

The written exam is administered monthly and the clinical exam by appointment (about six times per month). Students can challenge their results, using an established procedure, on either exam if they fail, but about half of those who challenge are still failed. The performance assessments do not evaluate all the skills required to be competent for entry into the occupation. For example, in

long-term care, only 3 skills are tested (selected randomly from 52 skills). In home health care, 13 skills out of 48 are tested. It takes 45 minutes to an hour to test one student and each student is tested individually. Performance evaluators are trained using guides developed by Oklahoma Vo–Tech Department. Evaluators are paid about $19 an hour to observe and score the tests. The quality of the multiple-choice portion of the exam is maintained in the same manner as the overall state competency-based system's quality.

AGC National Certification Administered by the Oklahoma State Vo-Tech Department

Description and Purpose

Because of Oklahoma's reputation for developing competency-based testing, the Associated General Contractors (AGC) organization hired the Oklahoma Vocational–Technical Department to develop assessments and administer a program that leads to nationally recognized credentials in three areas of the construction industry—carpentry (commercial and residential), bricklaying, and stone masonry. These are advanced certificates, with required prequalification of either two years of work experience or one year of work experience plus the completion of a vocational education program. Prequalification must be documented on the registration form before the test will be administered. Contracting occupations were included in the certification program, but occupations outside of AGC's "contracting" jurisdiction, such as plumbing and electrician work, were excluded.

AGC was incorporated in 1921 as a full-service construction association representing the needs of both open-shop and collective bargaining contractors. It represents 8,000 general contracting firms and 24,500 Associate and Affiliate Members; it has 101 chapters nationwide. Its mission statement states that "AGC is dedicated to providing programs that promote high standards in the construction industry. AGC has designed this certification program to give prestige and recognition to individuals working in the industry." AGC accredits training programs in various contracting trades. The association's members work mainly on commercial construction, where workers are most in demand. The incentive to sit for one of the certificates varies from chapter to chapter. AGC spends a lot of time teaching contractors that they need to invest in training the incoming work force.

The first AGC-sponsored tests were administered in 1989. Tests are multiple-choice, with high-level skills incorporated into test questions. Academic skills such as basic math and reading are included in the test. Oral testing is offered by special arrangement. They have never had a request for a test in a language other than English.

Implementation and Administration

Using the process developed for Oklahoma's vo-tech competency-based system, AGC's Workforce Development Committee oversees the certification program, including development of the task lists and multiple-choice tests. This committee and subcommittees in each certificate area are made up of contractors, training instructors, foremen, and supervisors. Tests are then administered through the 101 local AGC chapters across the country. Workers can be trained anywhere and then take the test. Curriculum materials have been developed and sold (on a voluntary basis) to various kinds of training programs (secondary and postsecondary vocational education programs, apprenticeship programs, and companies). This year 700 tests were scanned and 65–70 percent met the minimum passing score of 70 percent.

Task lists are reviewed and revised annually by committees of AGC contractors (the committees are coordinated by the state vocational-technical department). The committees select a pool of questions generated from the task list. These questions are entered into a test bank, which has grown over time. A test is then generated from the test bank every year. Tests range from 50 to 100 questions, depending on the tasks. Data on how important each task is to the particular job are taken into consideration in test development. Committees review tasks, test questions, and curricula annually.

In addition to developing curriculum materials and administering the test bank through an AGC subcontract, the Oklahoma Vo–Tech Department administers the tests through local chapters, scores completed tests, and conducts item analysis for AGC certifications. The Department has worked with AGC for 25 years; the last five have been highly focused on these certifications. AGC funds both a program coordinator (Pam Stacy) and a secretary at the department to run the testing program. AGC also employs a full-time curriculum developer at the curriculum center in Oklahoma. Oklahoma instructors can buy materials at cost, whereas AGC receives profits from sales in other states. The Vo–Tech Department is responsible for marketing the curriculum materials and tests.

The annual testing process begins in September, when chapters are asked to identify cities for test sites that year. About a third of the chapters request participation. In January, a box of promotional materials is sent to the chapters to advertise certification. Registration/test administration is $15 per person, per test (an individual may take more than one test). The test is then administered in April. Ms. Stacy works with test coordinators at each chapter to set up the test site and hire test examiners. The required conditions for each test site (such as lighting) and test examiners (such as a résumé demonstrating their work in a construction occupation) are specified.

Scores are reported for regions and test sites, if there are enough test takers. Individual results are sent only to the individual test taker. Sometimes Ms. Stacy is put in the middle between the employer and the employee. The employer may have paid for the test and wants to know the results, but they are confidential. Ms. Stacy compiles a report for each chapter with tips on how to improve scores next year.

Seventy percent is the minimum passing score. Research was done on the relationship of test scores to skill level and job performance, and the analysts decided somewhat informally that 70 percent was a "good" score for predicting successful job performance. However, there was no scientific research underlying this cut-off score.

Technical Quality

The AGC certification system is held to the same quality standards as the Oklahoma competency system. Content validity is high because tests are closely linked to the competencies developed by industry. Ms. Stacy says they would like to do more "concurrent" validity checks to look at correlations between the score and performance. After the test has been administered, through item analysis, she flags the questions that look problematic nationally. Committees make a decision about those questions, and scores are adjusted before they are reported to individuals. Ms. Stacy has a background in statistics, research, and psychometrics.

Consequences and Use of Assessment Results

Certificates can be used to document advanced training for raises and promotions, as well as hiring. AGC includes both unionized and nonunionized contractors. Union contractors utilize the apprenticeship system, whereas nonunion contractors operate an "open-shop," where employment is open to all regardless of qualifications. If a contractor hires only union members, certification can be used as an added qualification to help contractors decide which employees to hire; in some cases, employers are required to pay certified workers more. If it is an open shop, it is up to the contractor to decide how to use the certification. On average, 75% of the testers' fees are paid by employers.

AGC provides successful completers of the certificate with a hard-hat decal, wall certificate, and pocket card that prove they have certification. AGC tries to give the certificate prestige for both the workers and the contractors in order to encourage initial and advanced training. The association hopes that the certificates improve the performance of workers and build pride in skilled craftsmanship.

Applicability to Vocational Education

AGC aims to break even financially, with registration and test fees balancing the costs of the testing program, which means that most test sites need to recruit more test takers (with a limit of 20 test takers per site). Recruiting more test takers may be difficult in some areas, because of the limited incentives to take the test. When the program originally began, the committee thought that the $15 fee would allow the program to break even, or even make a profit. Despite having lost money on the assessment program, they have not considered raising the fee because they want to keep the tests affordable. Originally, there was national union resistance, so the program was more expensive and difficult to implement than anticipated. Currently, AGC has about 40 test sites in 25 states.

Because of the expense, performance assessments have not been pursued. Based on limited research, the advisory committee and state staff believe that multiple-choice test scores are highly correlated with job performance. In order to set up this type of system, a test-form scanner and customized software are required, as well as the capacity to predict and avoid (and, when necessary, respond to) legal challenges. Case law suggests that test takers may be tested on only those tasks required to perform on the job. Worried about legal challenges and other costs, AGC has steered clear of developing or implementing performance assessments.

Summary of the Three Oklahoma Assessments

Applicability to Vocational Education

Over a period of ten years, the Oklahoma competency-based assessments have developed into a system that has a good reputation and continues to branch into new areas—specifically, the AGC certification and the statewide health liscensure. The administrative and quality control procedures for multiple-choice examinations are already in place, and this system has conducted some experimentation with performance-based assessment. Although Oklahoma Vo–Tech believes that performance-based assessment is not a viable option statewide (for financial and logistical reasons), they are having some success with their statewide health licensing system, which includes performance tasks. However, reliability and consistency are still issues in all three of these assessments.

All three assessment systems discussed are applicable to vocational education, although with slightly different objectives and incentives for participation. The Oklahoma competency-based assessment system started as a way to certify students' skills and knowledge as well as move curriculum toward teaching skills and encourage instructional methods to focus more on having students

demonstrate competencies. However, many programs did not participate in the system. With the implementation of performance measures and standards, local sites are required to use the assessment system to report performance to the state and to make progress in program improvement. Many local instructors now see the system as "another state requirement" for accountability rather than as a certification system for helping students.

Oklahoma education officials see a need to certify not only students' occupationally specific skills with their current system, but also broader, more general skills through a career passport (a career-focused portfolio), which students will complete and present to potential employers. Once the passport system is implemented, state agency staff hope that educators will see passports and the vocational assessments as complementary parts of a certifying credential that students use to gain employment. It is hoped that the passport system will be viewed by instructors as integral to their curriculum. The six required components of the passport include:

- Documentation of educational skills (diploma);
- Documentation of vocational program completion;
- Documentation of competency in at least one occupation;
- Evidence of academic preparation that supports success in the workplace and eligibility to continue education;
- Evidence of meeting local attendance requirements; and
- Completion of a resume.

In the current system, the three health certificates (and three more under development) have had the most success, in combining multiple-choice and performance-based assessment, in breaking even on administration costs, and in recruiting test takers. The health certifications are required for liscensure, which are necessary for employment. Employers also need certified employees, and demand in these fields has been steady, especially in home-care occupations.

The AGC program was developed to provide advanced certification (not entry-level employment) in three areas in the construction industry. At this point, the program is not recruiting enough test takers to break even. The incentives for participation are based on demand and the strength of unions (and unions' preferred uses for the certification), which vary by locale. Therefore, participation rates differ considerably. The occupational areas being certified are not licensed professions and, therefore, participation is left to the discretion of employers, based on market conditions and observed employee performance.

Employers that hire both union and nonunion employees must value the certificate and incorporate it into hiring practices and salary scales for AGC to increase participation.

One process for developing an assessment has been adapted to develop the other two systems. The three systems have slightly different objectives and participants, but all are a part of the vocational education enterprise. For the most part, what we can learn from these cases is how to administer, nationally or statewide, criterion-referenced, competency-based, multiple-choice tests, with more limited lessons on locally designed and administered performance-based assessment. In many ways, vocational education in Oklahoma is atypical. The state strongly supports vocational education, with a large state staff and substantial funding. They operate with an entrepreneurial spirit that is not common in state bureaucracies. Consistent leadership and staff, university support, and a tradition of strong vocational education programs statewide have worked in their favor. One or more of these elements is likely to be absent in other states or metropolitan areas that may try to adopt this system or develop a similar one, so obstacles may arise.

D. The National Board for Professional Teaching Standards

The National Board for Professional Teaching Standards (NBPTS) was established to develop and administer a voluntary national certification system to recognize highly accomplished K–12 teachers. The standards and tasks by which candidates are judged were developed by groups composed primarily of other teachers. Once each assessment is fully tested and implemented, certification will be offered at 4 levels and in 14 subjects (e.g., one teacher candidate may apply for certification in high school English, or Early Adolescence/Math). In 1995-96, National Board certification was available in two categories: Early Childhood/Generalist (EC/G) and Middle Childhood/Generalist (MC/G).

To obtain the NBPTS certificate, teachers prepare an extensive portfolio demonstrating their preparation, classroom work, teaching strategies, and professional activities. In addition, they participate in two days of performance activities at a regional assessment center. The process takes about one school year to complete.

NBPTS certification offers benefits to teachers, school districts, and teacher training institutions. Teachers have an opportunity to reflect on and perhaps improve their teaching skills and professional life. School districts have an independent standard against which to measure the ability of their experienced teachers, and the process clarifies for teacher training institutions what accomplished teachers should know and be able to do. However, the ultimate, and most important, beneficiaries of improved teaching practices are the students.

Description and Purpose

The NBPTS, a nonprofit organization, was founded and initially funded through the Carnegie Corporation of New York to provide avenues for teachers to demonstrate their professional achievement. Establishment of the NBPTS in October 1987 fulfilled a major recommendation of the 1986 report *A Nation Prepared: Teachers for the 21st Century*, issued by the Carnegie Task Force on Teaching as a Profession. The Board hopes to improve the public's perception of

teachers, enhance teachers' own view of their profession, and consequently attract and retain high-quality teachers, all with the underlying goal of improving student learning.

Two-thirds of the 63 Board members are classroom teachers actively engaged in instruction, and the other third are public officials and others involved in education (e.g., governors, legislators, chief state school officers, board of education members, principals, superintendents, college presidents, deans, higher education faculty, parents, minority student rights advocates, and business leaders). A majority of the nonteacher Board members are elected or appointed public officials.

The Board has a threefold mission: "To establish high and rigorous standards for what teachers should know and be able to do, to certify teachers who meet those standards, and to advance other education reforms for the purpose of improving student learning in American schools" (NBPTS 1991). Its core activity is an assessment system organized around subjects and age/grade levels. Once the assessment system is fully implemented, teachers will be able to choose certification from among the following levels and subject areas (levels are combined for some subjects).

Levels:

- Early Childhood (Ages 3–8)
- Middle Childhood (Ages 7–12)
- Early Adolescence (Ages 11–15)
- Adolescence and Young Adulthood (Ages 14–18+)

Subject areas:

Generalist, English language arts, math, social studies/history, science, foreign language, art, music, vocational education, exceptional needs/generalist, English as a new language, health/physical education, library/media, and guidance counseling. While only EC/G and MC/G certifications were offered in 1995–96, the number will expand to six certifications in 1996–97.

Relationship to Other Programs

NBPTS certification complements but does not replace state licensing. State licensing systems set compulsory minimum standards for novice teachers, while the National Board Certification creates voluntary standards for experienced

teachers. Similarly, the National Board's standards should relate to but not substitute for requirements for preservice training. Teacher training institutions develop curriculum to comply with state laws, while the NBPTS standards establish a set of profession-endorsed guidelines regarding best practices schools can use in improving inservice training, for example when developing curriculum to train teachers.

To date, few states or districts recognize or reward NBPTS certification directly. However, some support it indirectly by paying teachers' costs for the testing process. One measure of the Board's success in creating a viable certification system will be an increase in the direct rewards to teachers who successfully complete the process.

The teaching profession is attempting to link accreditation, liscensure, and advance certification with the goal of ensuring that all students are taught by competent, professional teachers (Rahn 1995). Three national organizations, the National Council for Accreditation of Teacher Education (NCATE), the Interstate New Teacher Assessment and Support Consortium (INTASC), and the National Board for Professional Teaching Standards (NBPTS) are working with the two teachers' unions (the National Education Association and the American Federation of Teachers) to improve the profession of teaching for both teachers and students. They envision a linked system of preservice preparation, extended clinical training, and continuing professional development in which National Board certification plays an important role (see figure 1). Although much of this linkage is still under development, both NCATE and INTASC endorse the standards from which National Board certifications are built, and are taking steps to see that their own components of this process are aligned with the work of the Board. Similarly, relevant professional committees and other stakeholders are actively sought in every phase of implementing the system, from the composition of the Board, the Standards Committees, and the Field Test Network to the broad-based review of documents and test packages.

Implementation and Administration

Implementation of the teacher assessment system follows a process that first established a sound theoretical base and since then has continued to seek broad support among established educational organizations and stakeholders. The first task of the NBPTS was "to identify the knowledge, skills, and dispositions that describe accomplished teaching and to convert those attributes into high and rigorous standards upon which to base the National Board Certification system" (Strategic Plan for the NBPTS, no date, 1989?). Board staff reviewed the relevant

literature on these issues and the standard-setting work of other occupations. Expert consultants worked with the Board and its staff on key issues. Drafts of the report were widely circulated to leaders in the education community for their comments. The final report, *Toward High and Rigorous Standards for the Teaching Profession*, was adopted by the Board in 1989.

This document established the philosophical underpinnings for the program, including explaining the need for a national certificate program; prerequisites for applying (three years of teaching and, at a minimum, a baccalaureate degree); five propositions that set forth broad principles to guide the development of standards; and assessment activity development guidelines. The implementation timeline anticipated five years of research and development to specify standards in each field and to develop assessment products and delivery systems. But now, after eight years, only two certificates are available.

The NBPTS established a comprehensive organizational structure and process to develop the assessment system, with teachers playing a major role in almost every area (NBPTS 1991a). Standards committees with chairs appointed by the National Board have been established for each certificate field. Through a national competitive merit review RFP process, contractors were selected for several development and implementation tasks. Assessment Development Laboratories (ADLs) develop and pilot the assessment tasks. A School District Field Test Network (FTN) tests the assessment packages, providing candidates, administrators, scorers, and evaluations of the methods and systems. The test packages will be constructed by the Production Assembly Group. The Technical Analysis Group (TAG) provides research support to the other contractors (e.g., a literature review of assessment methods in other professions; developing the sampling frame of teachers for the field test trials) and synthesizes the work of the standards committees, ADLs, and other contractors. Additional research is independently contracted for.

The development process was implemented gradually. Initially, four standards committees were appointed, and then additional committees were added each year. Once the initial standards committees began work, an ADL was appointed to work with each of the two committees. A year and a half later, the RFP for six additional ADLs was issued. Slowly snowballing the organizational elements allowed the Board to learn from early experiences and adjust the process.

This extensive development process places the emphasis on teachers, while taking into account the knowledge and opinions of experts in relevant fields and other interested parties. Reaching consensus among such a broad stakeholder group, however, is a long and expensive process. Close to $57 million has been

spent to date, and the Board is concerned about making development and administration of the remaining assessments more efficient and economical.

Because the process of establishing the assessment system may provide a useful model for establishing a system for recognizing accomplished vocational teachers, it is described in considerable detail here.

Developing Subject Matter Standards. Standards committees are composed primarily of teachers but also include researchers and others involved in education policy and practice. The first four committees were appointed in 1990 for Early Adolescence/English Language Arts (EA/ELA); Early Adolescence/Generalist (EA/G); Adolescence and Young Adulthood/Mathematics; and Young Adulthood/Art. By 1995, a total of 17 committees were established to set standards in 21 of the more than 30 certification fields (Bradley 1995a), and by 1996 initial development had been completed and standards were released for public comment in all 21 areas.

Committee members are selected from a slate nominated by a broad group of professional organizations, Board members, and Board staff. For example, over 130 nominations were reviewed for the EA/G committee, and 12 members were chosen who provided a balance of gender, location, profession, and major focus (Hattie et al. 1994). Members of the relevant ADL are expected to participate in standards committee meetings. Liaisons from key professional organizations are also invited to participate as nonvoting members of the committees, e.g., the International Reading Association provided a liaison to the EA/G committee.

These committees develop draft standards for the knowledge and skills that teachers should have. Draft standards are reviewed by the Board, its Certification Standards Working Group (CSWG), the ADL, and a broad spectrum of stakeholders. Development of the EA/G standards took about three and a half years. A report from the EA/G committee (Hattie et al. 1994) notes the many iterations of standards, the continuous interplay with the Board and its staff, and later with the ADL; among the issues that arose was the difficulty of determining appropriate subject matter for generalist teachers. The report also emphasizes the final authority of the standards committee in making tough decisions.

Drafts of the standards were circulated widely. Members of the standards committee, field site members, and others who had contact with the committee were asked to rate the standards for clarity, for their relevance to highly accomplished teaching practice and for other factors. A four-month review period was established for comments from professional organizations, state bodies, teachers, academics, and other stakeholders, and the standards were revised in accordance with the comments. Almost 30 months after their first

meeting, the Board and CSWG approved the standards for the 1993–94 field test. Revised standards were approved the following year for the 1994–95 field test. While standard-setting proved to be a lengthy process, responses from a survey of field site participants, members of the standards committee, and Board members regarding the validity of the EA/G standards indicated widespread approval for the standards among both teachers and nonteachers.[1]

Assessment Development Laboratories (ADL). Assessment Development Laboratories work with one or more standards committees to develop and produce an assessment package. They are selected through a competitive merit review Request For Proposal (RFP) process. Fundamental principles guiding the labs were outlined in the RFP. Assessments must meet the following criteria: professional credibility, public acceptability, legal defensibility, administrative feasibility, and economic affordability. Multiple forms of assessments were to be employed, and respondents were to consider how student learning as a measure of teacher effectiveness might be demonstrated. The following three components were expected to be included in the assessment procedures:

- documentation from the candidates' school site, e.g., observations,
- videotapes, and/or portfolios;
- assessment of the candidate's subject matter knowledge
- and knowledge of child development for the specific age group; and
- extended exercises over several days at an assessment center.

The first ADL contract was awarded in 1990 to the University of Pittsburgh School of Education in conjunction with the Connecticut State Department of Education and six other state departments of education to develop assessments for the EA/ELA certificate. The second contract was awarded to the Performance Assessment Laboratory at the University of Georgia for the Early Adolescence/Generalist Certificate.

Both labs developed similar assessment activities that could be closely integrated with class lessons, such as developing appropriate applications for a new classroom resource, recording actual classroom plans and activities and analyzing what occurred, or analyzing and evaluating samples of student writing. The assessments emphasize giving teachers an opportunity to show

[1]Of 175 forms returned, 139 were from teachers and 36 from nonteachers. The report does not indicate the response rate to the survey. About 87 percent of teachers and 88 percent of nonteachers responded "agree" or "strongly agree" that "each of the 11 standards describes a critical aspect of highly accomplished teaching practice within this field" (Hattie, et al. 1994).

what they know in an authentic context, rather than pinpointing what they do not know. For both certificates, teachers complete activities at the school site incorporating their documentation in a portfolio, and perform additional activities at the testing center. Table D.1 compares the skills and activities targeted by each lab (Bradley 1994).

Another way to get a feel for the assessments is to approach them from the viewpoint of a candidate. Figure D.1 describes one teacher's experience in preparing a portfolio for the EA/ELA certificate is drawn from an *Education Week* account that follows two teachers through the entire process of the first field test (Bradley 1994).

The first round of certification (of which Diane was a part) revealed strengths and weaknesses in the assessment center model. The first teachers received certification in January 1995. More than one-third (81 out of 289) EA/G candidates participating in the 1993–94 field test were certified. Successful EA/ELA candidates were certified that summer. The two certificates were offered again in 1994–95 to fee-paying candidates and about 200 candidates participated. Fourteen assessors were required to score the EA/ELA exercises. Problems with this scoring resulted in a costly redesign of the scoring system and lengthy delays in announcing the results (Bradley 1995a). Because of these high costs, this particular certification will not be offered in 1995–96. This assessment will be revised and offered again in 1996–97. Reducing the complexity of the process will be critical. Modifications will focus on making the assessments less burdensome to scorers and candidates. According to James R. Smith, the Board's senior vice president, the portfolios, for example, asked for more material than was necessary, and should be more focused (Bradley 1995a).

Assessment Administration. The first field tests indicated that candidates felt they didn't have enough time to prepare their portfolios, so the time will probably be lengthened. The 100 hours teachers spent assembling their portfolios was about twice the time the Board had anticipated (Bradley 1994).

Findings from the administration of the first set of assessments (Scriven 1994) bring up several problems. Many candidates said that they were influenced to participate by the absence of a fee, thus the cost of the actual examination may discourage potential candidates. The description of the process needs to be clearer so that candidates know what to expect about the amount of time involved and the content of the exercises. About half of the candidates found the instructions for the portfolio exercises unsatisfactory, and most candidates felt the support provided was inadequate. Teachers sought more specific direction about activities, including, for example, the expected length for written

Table D.1

Comparison of Skills and Activities Targeted by Each Lab

Early Adolescent/Generalist	Early Adolescent/English Language Arts
School Site Activities	**School Site Activities**
1. Professional development and service: submit vita; write accounts of 1) an impact of professional development on practice and 2) professional service activity; obtain letters of support from colleagues.	1. Professional background: submit resume; write one- to two-page description of participating in a learning community.
2. Teaching and learning: write narrative describing a selected class over a period of time; describe the progress of three students, reflecting different learning characteristics; videotape class activities; and provide samples of student work and teaching practices.	2. Teaching and learning: describe and analyze the writing of three students, including the influence of instruction (submit with five to eight samples of the students' writing.)
3. Lesson analysis: select an unedited 30–45 minute videotape from a class and write account of the teaching and learning that occurred, highlighting five to seven particularly important points.	3. Interpretive discussion: videotape 15 to 20 minutes of a class discussing a piece of literature; write an evaluation of the discussion.
	4. Planning and teaching: write eight-page commentary describing planning and instruction over a three-week period, using an integrated curriculum that demonstrates cultural awareness; include a videotape of one class session.

Early Adolescent/Generalist	Early Adolescent/English Language Arts
Assessment Center Activities	**Assessment Center Activities**
1. Instructional resources: write analysis of the potential of SimCity for teaching social studies, math, history, and science. (SimCity is a computer simulation supplied to the candidate at the school site.)	1. Group discussion: With other candidates, develop a curriculum unit on personal relationships. Unit materials are selected from eight novels provided previously at school site. Discussion is videotaped.
2. Instructional analysis: write analysis of videotape and materials from a mathematics instructor, including suggestions for more effective strategies and extension of the topic to the arts.	2. Instructional analysis: analyze videotape of teacher-led discussion, including suggestions for improving instruction; show knowledge of young adolescent learning, and demonstrate cultural awareness and understanding of discussion dynamics.
3. Curriculum issues: After group discussion of a theme related to exploration of governmental systems, ecosystems, and the media, complete two-hour written description of the instructional development of a theme drawing on one of the above subjects.	3. Analysis of student writing: Analyze set of student papers and discuss analysis with interviewer, making suggestions for improving students' writing. Videotaped.
4. Content knowledge: three one-hour written subject examinations.	4. Content knowledge: three two-hour essay assessments on composition, literature, and language. Literature and journal articles are used for the essay prompts.

One Teacher's Experience

Diane Hughart is a middle school English teacher in a Virginia suburb of Washington, DC. She decided to volunteer for the EA/ELA assessment because she was curious: "I like the idea of trying something new and being involved." Rick, another teacher at her school, chose to apply for NBPTS certification, selecting the EA/G category. The two teachers worked together to interpret instructions, determine what to include in their portfolios, and buoy each other's spirits during the intensive preparation. The teachers also attended support meetings with other Fairfax County candidates and received some assistance from faculty at George Washington University's School of Education.

With only two months to prepare her portfolio, Diane first concentrated on choosing a class whose work she would document for three weeks. She settled on her sixth-period 8th-graders, a diverse group of outgoing students. Knowing that one component of the documentation was a videotape of the selected class, Diane quickly began videotaping class sessions, using students to do the filming. In this way she let the class get used to the camera so that she could capture an exemplary lesson on tape. Diane spent about 20 minutes each day filling out activity charts for that class. Her notes would culminate in an eight-page commentary on her teaching practices. Diane felt constrained by the chart format. She complained, "This is really very bland to me. It's humdrum. There's nothing in here about how I make my decisions. To me, reflection is what makes individual teachers different." Her colleague Rick, on the other hand, was required to complete a daily commentary on his selected class. He spent about two hours every evening writing the two- to three-page commentary and found the process very stimulating.

The preparation process occurred from mid-November through mid-January, a hectic time at best for Diane, with Thanksgiving and Christmas to plan with her two young adolescent sons. This year she also had sold her house and would be moving during the Christmas holidays. Renovations at school meant that her classroom would also be moved at about the same time. By mid-December she was feeling quite stressed. Students had not responded well to the videotaping. She found their class discussions "inhibited or silly." One particularly good lesson was interrupted by a fire drill. Since Diane routinely used portfolios to evaluate her students' work, she was not as concerned about assembling and analyzing writing samples from three students. But the hours she had spent documenting her class work put her behind on grading and returning student papers. Diane looked through many students' writing to pick three diverse students. Reviewing so much student work made her feel "more secure" about her teaching methods.

Diane chose a unit that used a play based on *The Diary of Anne Frank* to show the integration of reading, speaking, writing, listening, and viewing. Students read the play aloud, discussed it in groups, watched a film version, kept their own diaries, and wrote about their own holiday traditions. At a support meeting, she told the other teachers, "I am past the stage of trying for perfection, I am just trying to get it done." The weekend before the portfolio's Friday deadline, Diane selected a video to submit; meanwhile, her dining room table was covered with stacks of materials for her portfolio. The last two nights (until 3 A.M. and 2 A.M. respectively) were spent putting the materials in order. The directions had been intimidating and Diane was disappointed in herself. "I thought I would be proud of it," she says, "but I'm not. I feel like it's not good enough." Diane finally arrived at the post office at 11:00 P.M. After a few days' rest and reflection, Diane felt more positive about her portfolio and what she learned from looking at her students' work and meeting new people. In early March 1994, Diane and Rick went to the assessment center for two more days of activities. A happy postscript: in June of 1995, Diane learned she was among the first group of teachers to earn NBPTS certification in the EA/ELA category.

Figure B.1—One Teacher's Experience

assignments.[2] Peer support groups were judged very successful in helping to prepare materials; however, the help of principals was not useful. About half of the participants found preparation workshops and video support useful.

At the testing center, about three quarters of the candidates bemoaned the lack of computers for the writing tasks; the amount of writing required was also judged excessive. Observers also indicated that testing coordinators needed to be better trained. The original 12-hour day was problematic; it was subsequently reduced to 8 hours. A particularly troubling finding was that about 40 percent of participants felt that seeking certification placed them at some risk in their schools. The evaluator of the test administration (Scriven 1994) noted that this was "consistent with other evidence that teachers tend to identify efforts to excel as egotistical or undemocratic." Scriven warned that if merit pay was tied to certification, it could increase negative reactions, particularly if principals share this attitude.

Based on the experiences of the first two labs, the draft RFP for subsequent ADLs outlined a streamlined process that would save time, money, and human resources (NBPTS 1991c [draft RFP]). The following six assessment methods were specified: (1) a portfolio of classroom teaching accomplishments including evidence that the candidate participates in a learning community, samples of student work, and artifacts produced by the teacher, (2) observations of teachers in their classrooms, (3) structured interviews based in part on the portfolio, (4) exercises typical of teachers' work, e.g., viewing a video of teaching situation and grading samples of students' work resulting from the situation, (5) simulations that are "contextual assessments," e.g., suggesting more effective strategies after viewing a videotape of a teacher's performance, and (6) written tests of subject matter knowledge and pedagogy.

Contractors were required to develop assessment methods for four teaching fields "with exercises and shells that cut across fields where appropriate." At the school site, teachers are asked to assemble portfolios of their teaching practices, including evidence of participation in a learning community and examples of student work; and they undergo multiple observations. At the assessment center, methods include structured interviews, simulations, assessments of subject matter knowledge and pedagogy, and exercises that assess skills critical across teaching fields, e.g., monitoring students' learning.

---

[2] An *Education Week* article (Bradley 1994) notes that among three Fairfax County, VA, candidates for the EA/G, one's teaching and learning commentary was 66 pages, another's was six pages, and a third's was one page.

Field Test Network (FTN). Once the labs have developed tasks, they are field-tested through a national Field Test Network of more than 100 school districts. The network includes 165,000 teachers, 25 percent of whom are members of a minority group. Districts in the network have a two-year contract to perform a variety of tasks, including reviewing standards, developing staff development programs for candidates, and field-testing assessment packages (NBPTS Press Release 6/22/92). The sites participate on an as-needed basis as the examinations are developed. For example, 26 sites participated in the first field test.

Scoring Assessments. Scorers are recruited nationally and receive training in the goals and standards as well as specific tasks. Exercises are scored independently by at least two teachers who themselves met the criteria for the relevant certificate during the first round (no one had yet received formal certification). If there are differences in the scores, the scorers meet and discuss the evidence, then independently rescore the exercise. If differences still exist, a third scorer is brought in. The final score represents agreement between two of the three scorers (NBPTS 1995).

Technical Quality

Since it began to develop its certification system, the NBPTS has been concerned about producing a high-quality and technically defensible process. It issued a Request for Proposals to establish a Technical Analysis Group (TAG) of researchers and psychometricians who could offer guidance to the assessment developers and evaluate the quality of the certification procedures. During its first year, the TAG either conducted or commissioned eight studies of the technical quality of the 1993–94 EA/G assessment.[3] (A more detailed description of each of the eight studies follows at the end of this case study report.) The topics of these studies were:

[3]The eight studies are: (1) "A Description and Evaluation of the NBPTS' Initial Process for Establishing Teacher Certification Standards," John Hattie, et al.; (2) "Matching Exercises, Aspect guides, and Decision Guides to Standards of the Early Adolescence/Generalist Certification Process: A Preliminary Content Validation," B. Loyd, et al.; (3) "Quality of Field Test Operations," M. Scriven; (4) "A Formative Evaluation of Scorer Training for Early Adolescence Generalist Exercises," D. Felker; (5) "A Commissioned Study of the Application of the Early Adolescence Generalist Scoring System to 1993–1994 Early Adolescence Generalist Candidate Submissions," C. Heider, et al.; (6) "Report on a Study of Decision Consistency Based on Data from the 1993–1994 Field Test of NBPTS's Early Adolescence Generalist Assessment," R. Traub; "Report on a Study of the Generalizability of Scores Earned on the Seven Exercises of NBPTS's Early Adolescence Generalist Assessment Based on Data from the 1993–1994 Field Test," R. Traub; (7) "Recommended Performance Standards for NBPTS's Early Adolescence Generalist Assessment," R. Jaeger, et al.; and (8) "An Analysis of Adverse Impact in Rates of Certification on the Early Adolescence/Generalist Assessment," L. Bond and R. Linn.

1. the development process for content standards,
2. content validity,
3. quality of field test operations,
4. quality of assessors' training,
5. validity of the application of scoring procedures,
6. consistency of certification decisions and reliability of exercises,
7. recommended performance standards, and
8. adverse impacts of differing certification rates of diverse groups.

A panel of respected educational researchers was convened to review the development of the certification system and determine whether the process was sound from a technical standpoint. They concluded that there were "no technical impediments to the Board's use of its Early Adolescence/Generalist assessment to award National Board Certification to candidates whose performances satisfy the . . . final recommended performance standard" (Bond et al. 1994).

However, the expert panel also recommended that the following issues be given further study:

- Increasing the reliability of the assessment (to resolve problems about scores near the passing standard);
- Whether having two content standards assessed less frequently than the others is acceptable;
- Exploring strategies to reduce possible adverse impact, i.e., the likelihood that the percentage of African-American candidates who would be certified was far lower than that of non-Hispanic white candidates;
- Developing additional forms of the assessment center exercises.

Use of the Results and Effects

Substantive Lessons from the Field Test. (Bradley 1995b). Assessment developers learned several lessons from the scoring exercises and debriefing of participants. Scoring of the first set of portfolios indicated that teachers were not skilled at reflecting on their own work; they were also more comfortable at describing than analyzing teaching practice. As a result, future assessments will include more focused open-ended questions. Moreover, interviewing teachers at the assessment center about the work in their portfolios was also problematic. Identifying and training skilled interviewers is difficult and costly, and this is not

the most effective assessment method. Classroom videotapes and samples of students' work proved to be more reliable measures of teacher practices than teachers' own descriptions.

While point-in-time samples of students' work provided little information about how students' learning progressed, they could provide worthwhile information about the value of teachers' assignments. Developers also learned the importance of framing activities to evoke the skill they wanted to evaluate, e.g., a videotaped presentation by students about a class project does not help assessors evaluate the teacher.

Because the program is in its infancy, little in the way of rewards has been implemented out in the field. Merit pay, mentor status, the right to teach in any state, and waiver of credential renewal requirements are all rewards that school districts or states may consider. Only North Carolina financially supports teachers who are pursuing the certificate and rewards teachers who obtain it. North Carolina pays the assessment fee, provides several days' preparation time, and a 4 percent salary increase for teachers who obtain NBPTS certification (Hunt 1995). Iowa, North Carolina, New Mexico, and Oklahoma waive state licensing requirements for NBPTS-certified teachers who move to the state. Massachusetts and Ohio accept the NBPTS certificate in lieu of their own state recertification (Richardson 1995).

The NBPTS also recognizes that the assessment process itself may give teachers an opportunity to grow professionally by reflecting on their skills and knowledge, and having the opportunity to measure themselves against objective, peer-developed standards (NBPTS 1995). Test developers also hoped that exposure to new techniques such as computer simulations may provide professional growth opportunities, e.g., by exploring the relevance of a computer simulation to their instructional strategies (Capie et al. 1995). One of the EA/ELA candidates who participated in the initial field tests (Bradley 1994), commented that the whole experience felt well integrated. "The process, she notes, taught her some things that she plans to incorporate into her teaching. But, she adds, it wasn't a 'major, earth-shattering event' in her professional life" (Bradley 1994). The EA/G candidate found that "examining his own professional development was a worthwhile activity that helped him clarify who he is as a teacher. Exploring ways to integrate other subjects into his lessons was particularly exciting."

Applicability to Vocational Education

The NBPTS's long-term goals for increasing the professionalization of teaching have not yet been realized, although much progress has been made. Key steps for developing the certification system have been undertaken, and many have been completed. For example, standards committees have been formed, and most have completed their initial work. Draft standards in 21 of 30 areas have been released for public comment. On the other hand, progress in implementing the assessments has been slow. Certification was offered in only two areas in 1994–95, and only six certificates, instead of the projected nine, will be available in 1996–97:

- Early Adolescence/Generalist
- Early Adolescence/English Language Arts
- Early Childhood/Generalist
- Middle Childhood/Generalist
- Adolescence and Young Adulthood/Mathematics
- Early Adolescence through Young Adulthood/Art

As anticipated, the cost of developing the NBPTS assessment system was high. Unfortunately, both the time required for development and the cost of administration were greater than anticipated. The NBPTS has received over $50 million since its inception in October 1987—$37 million from private donors and foundations and $19.34 million in one-to-one matching funds from the federal government (Bradley 1995a), but this has not been enough money to maintain the initial development schedule. Recently, cost overruns have caused the organization to cancel three of the seven ADL contracts. The desire to produce innovative assessments (using "authentic" measures wherever possible) while maintaining high standards for technical quality has pushed costs upward.

Furthermore, assessment center activities are proving to be more costly than anticipated. For example, administering the evaluation to the first group of candidates cost $4,000 per teacher (not counting development costs). "Most of the expenses . . . went to scoring each candidate's work, an exhaustive process that in some cases took 23 hours." Costs for the second set of EA/G and EA/ELA tests were reduced to $3,000 per participant. And costs for the field tests of the Early Childhood/Generalist and Middle Childhood/Generalist certificates will be about $2,500 per participant (Bradley 1995a). The application fee of $975 does not come close to covering these costs.

Recently, NBPTS assembled a team of experts to offer suggestions for bringing down the evaluation costs. Two of the suggestions under consideration were reducing the testing center assessment to one day instead of two, and using the portfolio as a screening device to limit the number of teachers that would be invited to the assessment center (Bradley 1995a).

On the other hand, candidates who complete the process find it to be extremely rewarding. Yet the burden on candidates is substantial; some applicants report that they spent approximately 120 hours on the assessment process. Candidates for the first two field tests reported spending about 100 hours each assembling the portfolios (Bradley 1995a).

The National Board experience offers two kinds of lessons for vocational educators. First, even ignoring complex legal questions that must be considered when tests are used for employment-related purposes, the NBPTS experience shows how difficult it can be to use alternative assessments when high stakes are associated with the results. Candidates expect that National Board certification will be accompanied by professional recognition and even financial rewards. The Board intends for the certification process and standards to drive preservice and inservice training and affect state licensing standards. Consequently, they adopted an approach to insure that the standards were endorsed by the profession and the certification process met the highest criteria for quality and fairness. In the case of standards, that translated into multiple stakeholders, frequent review, and systematic quality control procedures. In the case of the assessments, the National Board's commitment to use alternative assessment strategies and to maintain high standards for the reliability and validity of decisions has made the process complex and necessitated additional professional review and analysis. All this translates into time and expense.

The experience of the Assessment Development Laboratories offers a useful picture of the complexity associated with alternative assessments. Following the Board's lead, the ADLs tried to be innovative and rely on "authentic" assessments to measure teacher competence. They developed interesting, relevant and meaningful exercises for teachers, but often found it difficult to score these exercises fairly and consistently. In some cases the activities did not necessarily reflect the underlying competencies they were designed to measure, and in other cases it was difficult for raters to agree on the quality of a candidate's performance. Because they were using new assessment methods, they often could not rely on traditional approaches to monitor quality, but had to improvise and rely on judgment. The Board employed a special Technical Analysis Group to consult with and review the work of the ADLs, because the developers were breaking new ground. In one case, the certification process was

delayed for months while a new scoring process was implemented to replace one that proved to be inadequate. This experience provides a sense of the level of difficulty that may be encountered if alternative assessments are used in high-stakes contexts.

These issues should not be unfamiliar to vocational educators, particularly those in the health professions. The requirements for professional credibility and legal defensibility have led to very thorough and comprehensive assessment efforts in the health fields. However, in most cases these professional assessments use traditional techniques, such as multiple-choice and short-answer questions. It is the combination of certification and alternative assessments that creates challenges for both quality and defensibility. Vocational educators have used alternative assessments at the classroom levels for years, because they link classroom and work experience more closely. The concerns raised here should not weigh too heavily on teachers who want to use more authentic assessments as part of their program. Quality issues demand greater attention if the assessments become part of a certification system that has important rewards for students.

The second lesson that can be drawn from the experience of the National Board relates to the qualifications of vocational educators. The Board intends to offer a single certification in Vocational Education (Early Adolescence through Young Adult). However, they encountered some difficulty in developing the standards because of disagreement about whether it was valid to use general standards for all vocational educators or whether each occupational area had to be treated separately. In the end, the standards development committee agreed on a single set of standards that is now available in draft form for public comment. However, the discussion of what general standards covering vocational education should look like is likely to increase in importance as education moves toward greater integration of vocational and academic curricula and as emphasis shifts in vocational programs from specific occupational skills to broader aspects of an industry.

Addendum:

Discussion of the Eight Studies of the Technical Analysis Group

The following sections describe the eight studies by the Technical Analysis Group (TAG), which explored the sequence of steps carried out by the National Board to develop the Early Adolescence/Generalist Examination.

Development of Content Standards. Hattie et al. (1994) examined the process for establishing content standards for the EA/G certificate. The study examined a number of questions, including:

- did the process for selecting committee members result in a diverse group of highly respected professionals;
- were adequate instructions given about roles and responsibilities;
- was the process fully documented; were comments from stakeholders (the Board, teachers, professional organizations, and experts in the field), the public, and experts in the field sought and attended to;
- did stakeholders have confidence in the standards that were established;
- and was the ADL included in the process.

The authors concluded that "a major initiative to ensure that a valid process for establishing standards has taken place to encourage input, criticism and reactions from the expert judgment of practitioners in the field." It is important to note that the study looked only at the *process* of establishing standards, not the content validity of the standards (the letter would ask how well the standards measure the knowledge base of an accomplished EA/G teacher). One important lesson is that the process to establish acceptable standards for one certificate category was quite lengthy, stretching over four and a half years.

Content Validity of the Assessment Tasks. Lloyd and Crocker (1994) examined the content validity of the EA/G tasks, using the members of the EA/G Standards Committee as expert panelists. After receiving training in content validation, the panelists analyzed each of the exercises and rated them on two factors: their relevance to each of the 11 standards developed by the EA/G committee, and the importance of the task to highly accomplished teaching. Among other steps, the panelists rated the guides given to the task assessors on their relevance to each of the 11 standards.

The researchers concluded that the exercises and decision guides were generally relevant to the EA/G standards; however, they cautioned that their evaluation was preliminary and should be followed up with a more comprehensive evaluation of content validity in the future. The majority on the panel did note that one standard was not addressed in the assessment exercises; other shortcomings were discussed as well. Finally, the authors pointed out that the EA/G Standards Committee had a vested interest in the certification procedure, and that no content validity study had been done on the standards themselves, so the parameters of accomplished EA/G teaching defined by NBPTS may be inaccurate.

Quality of Field Test Operations. To evaluate the administration of the exercises at the test centers, three ethnographers attended testing and scoring sessions at different test centers, focus groups were held with the directors of each field test center and with field test coordinators, candidates completed questionnaires at the conclusion of the session, focus groups and individual interviews were conducted with a sample of teachers at three sites and comments were collected from interviewers, proctors, dropouts, and from unsolicited persons. In a summary description, Scriven (1994) concluded that test administration is working well enough to justify continuing, but that marked improvement needs to occur in each of the next two years.

As might be expected, considerable logistical problems were evident, including travel reimbursement questions, addressing controls for cheating, timely and comprehensive supply of instructions and materials for training assessors, the need for backup staff, and provision for handicapped access. Candidates complained about the lack of word processors, excessive writing requirements, the 12-hour day, insufficient time for portfolio preparation, and the inadequacy of support. Questions that will become increasingly important are the face validity of the assessments, the potentially negative response to candidates from their school colleagues, and the cost of the assessment. Some of these issues are explored further in the section "Use of the Results."

Quality of Assessor Training. Felker (1994) reports on the training of the assessment scorers. Experienced professional educators and training specialists observed training sessions at all four training sites. Observers reviewed many aspects of the training sessions, including: recruiting scorers, training design, trainers, training materials, exemplars, scoring rubrics, training activities, qualifying process, duration, atmosphere, and replicability. Most sessions had one observer. Felker (1994) synthesized the reports and concluded that the greatest strength of the training was the quality of the trainers. Most had been involved in developing the exercises and were very knowledgeable both about their content and about how to score the complex candidate responses. They were also "sensitive and responsive instructors."

There was substantial doubt among observers about the ability to replicate this high-quality training in the future. The materials for training subsequent trainers were insufficient to duplicate the content knowledge and procedural knowledge existing in the initial group of trainers, so intensive training of them would be required. Other problem areas included the training materials given to scorer trainees (some needed editing, reformatting, better graphics), characteristics of selected scorers, and the qualifying process. However, the author believed that "the training process for EA/G was of sufficient quality to teach assessors how to score. Clearly, the training was not without problems and deficiencies, but the overall process was sound and people learned what was taught."

Validity of Scoring Procedures. A panel of 11 experts, 10 of them members of the task content validity panel, evaluated the scoring system (Heider et al. 1994). Panelists individually ranked pairs of candidates' responses to one of the assessment exercises and then compared their rankings with rankings assigned by the EA/G scorers. Panelists also analyzed candidates' responses to a given exercise and made three specific judgments: whether the assessor used a base of knowledge, and skills within the subject area assessed by the exercise; whether the assessor's judgments reflected the entire subject area for the standards being assessed in the exercise; and whether the assessor's application of the standards being assessed by the exercise was appropriate.

Panelists concluded "that the scoring procedures used on the candidate responses we analyzed demonstrate high levels of rank-order agreement and withinness" ("withinness" refers to whether the grounds used by the assessor were within the content domain assessed by the exercise). However, panelists were concerned that "not all standards were applied in a representative fashion and, in some cases, standards were not applied as rigorously as they might have been." Before another assessment is administered, the panel recommended that explicitly connected scoring guides be made with standards, so that candidates are not told they are being assessed on a given standard if that is not the case, and that assessors be trained to score with more precision. The panelists also recommended that conditions that might affect scoring be examined further, e.g., writing skill of the candidate, how well the candidate follows directions, and why candidates fail to address the specific task assigned.

Consistency of Certification Decisions and Reliability of Exercises. A quantitative analysis was conducted to determine whether consistent pass/fail decisions were made for each candidate (Traub 1994a; Traub 1994b). Traub found that "the training of scorers seems to have produced, for each exercise, a cadre of scorers who can achieve an acceptably high level of consistency in discriminating among candidate responses." He concluded that the scoring of the EA Generalist

exercises "was done with sufficient care that unreliability due to assessor differences should not be a matter of concern to the National Board in deciding whether or not to implement this assessment."

Recommended Performance Standards. Jaeger (1994) presents a description of the process used to develop the performance standards that are used to classify candidate performance into four levels: "Superb," "Accomplished," "Competent," or "Deficient." The report also provides information on the pass/fail rates that would result from various standard-setting rules. A panel of 30 teachers who had scored 3 or higher (on a 4-point scale) on 3 of the 5 exercises and 2 or higher on the remaining 2 exercises was assembled to develop the performance standards. Consideration also was given to achieving racial, ethnic, and gender diversity on the panel. Panelists received two days of instruction about the goals of the NBPTS, the specific exercises, and the scoring procedure.

Panelists developed scoring standards based on scores of 200 hypothetical candidates. If these standards had been applied to actual candidates, only 7 percent would have been certified. The panel than considered different scoring standards and eventually settled on standards that would mean 21 percent of candidates would be certified (including one African-American, one Hispanic, one Native American, and 60 white teachers, of whom 55 percent would be female). A final adjustment in the performance standards policy was made following a recommendation by a core group of the TAG. This adjustment raises the certification rate of the EA/G candidates to 27 percent.

Analysis of Adverse Impacts of Certification Rates. Throughout the development and administration of the NBPTS assessment, efforts have been made to reach diverse of representation on committees, panels, work groups, and among staff and candidates. Bond and Linn (1994) examined the impact of race, gender, and school location (urban, suburban, rural) on performance on the assessment exercises. No evidence of adverse impact was found for gender or the location of the school where candidates teach. However, "Substantial adverse impact with respect to race was found. Of the 40 African American teachers in the field test, 20 submitted complete, scoreable responses, and only one (5 percent) would gain National Board certification." This compares with 39 percent of white teacher candidates who would be certified. (The number of Hispanic, Asian American, and Native American teachers in the field sample was too small to reliably judge their relative performance.) Reliable differences between scores of African Americans and white candidates were found on all of the exercises. Analysis also determined that the difference in scores could not be attributed to the amount of writing required, similarity or difference between assessors' and candidates' race, or differences in perceived levels of support. Nor was there

material in the assessments that might be considered sexist, racist, or otherwise offensive to the EA/G teacher population. In sum, though adverse impact was noted, it could not be attributed to bias.

Bibliography

Bond, L., and R. Linn. (1994). "An Analysis of Adverse Impact in Rates of Certification on the Early Adolescence/Generalist Assessment." Greensboro, NC: Technical Analysis Group, National Board for Professional Teaching Standards.

Bond, Lloyd, L.J. Cronbach, E. Haertel, R.M. Jaeger, R.L. Linn, and B. Lloyd. (1994). "Conclusions on the Technical Measurement Quality of the National Board for Professional Teaching Standards' Early Adolescence Generalist Assessment." Greensboro, NC: Technical Analysis Group, National Board for Professional Teaching Standards.

Bradley, A. (1994). "Pioneers in Professionalism." Education Week, April 20, 1994.

Bradley, A. (1995a). "Overruns Spur Teacher Board to Alter Plans." Education Week, May 31, 1995.

Bradley, A. (1995b). "Teacher Board Providing Valuable Lessons in Using Portfolios." *Education Week*, May 31, 1995.

Cape, W., L. Dickey, and J. Anderson. (1995). "Design Considerations for NBPTS Assessments." Performance Assessment Laboratory, College of Education, Athens: University of Georgia (paper presented at the Annual Meeting of the American Education Research Association).

Felkner, D. (1994). "A Formative Evaluation of Scorer Training for Early Adolescence Generalist Exercises ." Greensboro, NC: Technical Analysis Group, National Board for Professional Teaching Standards.

Hattie, J., P. Sackett, and J. Millman. (1994). "A Description and Evaluation of the NBPTS' Initial Process for Establishing Teacher Certification Standards, Second Draft." Greensboro, NC: Technical Analysis Group, National Board for Professional Teaching Standards.

Heider, C., J. Herbert, J. Lashley, R. McLeod, J. Perry, D. Strahan. "A Commissioned Study of the Application of the Early Adolescence Generalist Scoring System to 1993-1994 Early Adolescence Generalist Candidate Submissions." Greensboro, NC: Technical Analysis Group, National Board for Professional Teaching Standards.

Hunt, J. B. (1995). "A Course That Creates A+ Teachers." USA Today, October 23, 1995, p. 13A.

Jaeger, R., 1994, "Recommended Performance Standards for the National Board for Professional Teaching Standards' Early Adolescence Generalist Assessment." Greensboro, NC: Technical Analysis Group, National Board for Professional Teaching Standards.

Lloyd, B. and L. Crocker. (No date, probably 1994). "Matching Exercises, Aspect Guides, and Decision Guides to Standards of the Early Adolescence/Generalist Certification Process: A Preliminary Content Validation." Greensboro, NC: Technical Analysis Group, National Board for Professional Teaching Standards.

Mandel, D. (No date). "Notes from a Meeting of the Technical Advisory Group." University of North Carolina, Greensboro.

National Board for Professional Teaching Standards. Executive Committee. (No date, January 1989?). "Strategic Plan for the National Board for Professional Teaching Standards."

National Board for Professional Teaching Standards. (1990a). "Request for Proposals for Award of Contract to Establish the First Assessment Development Laboratory for the Early Adolescence/English Language Arts Certificate." April 11, 1990.

National Board for Professional Teaching Standards. (1990b). "Request for Proposals for Award of Contract to Establish the Second Assessment Development Laboratory for the Early Adolescence/Generalist Certificate." July 16, 1990.

National Board for Professional Teaching Standards. (1991a). "Research and Development Plan."

National Board for Professional Teaching Standards. (1991b). "Toward High and Rigorous Standards for the Teaching Profession."

National Board for Professional Teaching Standards. (1991c). "Draft Request for Proposals for Multiple Assessment Development Laboratories, RFP #6, " 1991, NBPTS.

National Board for Professional Teaching Standards. (1992). "School Districts Join Nationwide Network in First Step Toward National Teacher Certification System." press release.

National Board for Professional Teaching Standards. (1994). "How We Plan to Achieve Our Vision."

National Board for Professional Teaching Standards, 1995, "An Invitation to National Board Certification."

Richardson, Lynda, "First 81 Teachers Qualify for National Certification," The New York Times, January 6, 1995, p. A1.

Scriven, Michael, 1994, "Summary Report on the Administration of the Assessment Process for NBPTS," Evaluation & Development Group, Inverness, CA. report to the Technical Analysis Group, National Board for Professional Teaching Standards.

Traub, R., 1994a, "Report on a Study of Decision Consistency Based on Data from the 1993-1994 Field Test of the National Board for Professional Teaching Standards' Early Adolescence Generalist Assessment," report to the Technical Analysis Group, National Board for Professional Teaching Standards.

Traub, R., 1994b, "Report on a Study of the Generalizability of Scores Earned on the Seven Exercises of the National Board for Professional Teaching Standards' Early Adolescence Generalist Assessment Based on Data from the 1993-1994 Field Test," report to the Technical Analysis Group, National Board for Professional Teaching Standards.

E. VICA National Conference: Job Skills Contests and Leadership Development Contests

Description and Purpose

Vocational/Industrial Clubs of America (VICA) is a national student organization for secondary and postsecondary students in vocational/technical fields. The occupationally oriented skills tests that form the centerpiece of VICA's national conference, called the United States Skill Olympics (or USSO), cover a broad range of vocational fields, test generic and job-specific skills, and use several different forms of assessment. VICA, a nonprofit organization, has gained assistance from corporate leaders and practitioners in some 60 fields to develop the tests, which are designed to measure skills required in that field anywhere in the nation. The two main divisions of these contests are the *Job Skills contests*, in which individuals compete in performing job-related skills and applying relevant knowledge from the vocational field they are studying, and the *Leadership Development contests*, which consist mainly of demonstrating generic and employment-readiness skills (some also require using academic skills, such as the contests in speech-making or parliamentary procedure). The national competitions are the culmination of local, regional, and state contests; winners proceed to the next level.

Students, teachers, and industry professionals who serve as Skills contest competitors, judges, technical advisors, education team members (for liaison and communication), and advisors/chaperones were in general quite positive and enthusiastic about the VICA program and the system of Skills contests, including local, state, and national competitions. While it is not surprising that students who have won local and state competitions would be excited about attending the national conference, it is more telling that high-level managers and other employees in the industries VICA serves were strong supporters of the program. Many have been involved year after year in designing the tests, working out logistics, or judging the competitions, and many also were taking personal time off from their jobs to do this work (entailing several days to a week of very long hours). In order to hear the down side of VICA's programs, one may have to locate and interview people who had participated earlier in the process but were eliminated before the nationals, or left due to disagreement with organizational policies or practices.

In some cases this dedication may reflect appreciation for VICA's having helped them at the start of their careers, but it also attests to positive impressions of VICA's ongoing work; it seems unlikely that these professionals would continue to work on the conference and persuade their companies to donate equipment and materials if they weren't consistently recruiting good employees through VICA. Industry representatives, mainly managers or laborers in the skilled trades, are complemented by industry association employees, vocational teachers (secondary and postsecondary), labor union representatives, and vocational administrators (school or local/state agency) in putting together this wide-ranging conference.

The VICA assessments focus mainly on demonstrating hands-on occupationally specific skills, though many also call upon students to use cluster-focused or more general industrywide knowledge (such as fundamentals of electronics, in the electronic products servicing contest, or the properties of particular metals, in precision machining technology) or generic thinking, decision-making, and trouble-shooting/diagnostic skills (e.g., in automotive repair technology or residential wiring) as well as specific academic skills (e.g., math skills in cabinetmaking and automated manufacturing technologies).

In many of the contest areas, a written exam is included; items are often multiple-choice but sometimes constructed-response. The different components are designed to test the skills and knowledge a person needs to work in a specific occupation or occupational group, so the parts are intended to complement one another. (See more detailed discussions of several Skills tests, in Figure E.1.) The results are used to determine first-, second-, and third-place winners in each competition among, first, secondary students and, second, postsecondary students; they're ranked in two separate categories, although in most contests they are assigned the same tasks, with some rare exceptions. The winners are determined by the total scores on all components, or contest stations, from all judges. Many times, the scores are so close that a fraction of one point (out of a score typically in three digits) determines the winners. The top three scorers in each category (secondary and postsecondary) win medals along with prizes like scholarships or equipment (other participants are given certificates). The first-place winner in certain skill areas (those included in the international competition that VICA participates in) competes in a runoff against the winner from the previous or next year, and the runoff winner receives additional in-depth industry-sponsored training to prepare for the international competition.

The questions and projects are judged using criterion referencing. In most contests, judging combines objective checks against a single right answer (a multiple-choice, yes/no, or mathematical question, for example) with aesthetic or

The vocational-technical areas represented in the Skills contests range from trade and industrial crafts (e.g., auto mechanics, brick masonry, carpentry, printing) to home economics–related or service occupations (culinary arts, commercial sewing, cosmetology, practical nursing, advertising design) to emerging or rapidly changing technology-based occupations (automated manufacturing technology, robotic workcell technology). Some assessments in Leadership Development pit teams of students against each other, but most assessments require strictly individual performances. Below are brief descriptions of the specific tasks assigned for a few of the assessments.

Electronic Products Servicing

The contest consists of three main parts: assembling an electronic product following a schematic drawing, diagnosing the malfunctioning component in several products and identifying an appropriate repair strategy, completing a written exam on electronics facts and theory, and demonstrating safety procedures throughout. Associated skills that are needed to perform these tasks include selecting appropriate test equipment, following safety procedures, soldering and desoldering, and performing tasks quickly (as well as correctly).

Law Enforcement

A written test covers constitutional and criminal law, main principles of the U.S. criminal justice system, rules of evidence, the law enforcement code of ethics, and similar topics. Contestants must also respond to video-supported scenarios involving, for example, an armed robbery in progress (including facing the split-second decision on whether to shoot at suspects); follow proper procedures in conducting an initial investigation of a threatening situation, make an arrest, collect evidence, and fill out an evidence collection form.

Culinary Arts

There are two separate contests, one for secondary and the other for postsecondary (the latter is more difficult), but in each students must prepare several platters of cold foods and a multicourse meal of hot/cooked items using ingredients, equipment, and tools provided. Judging covers the following elements: sanitation and safety; *mise en place* (visual presentation), organizational skill, technical skills (like chopping, slicing, sautéing, kneading dough), quality of prepared items (taste and smell), and creativity.

Precision Machining Technology

A range of projects includes: turning or milling a piece of aluminum on a lathe to specifications in a blueprint (two separate tasks); interpreting a blueprint and answering questions, technical sketching; and benchwork, including layout, deburring, assembly, filing, drilling, grinding, hacksawing, and fitting; making calculations using gauge blocks; and implementing precision measurement (micrometer variations and transfer measurement).

Figure E.1a—Some Specific Contests

Advertising Design/Commercial Art

There are three main components, in which students must 1) compose a camera-ready mechanical (pasteup) according to specs for a print ad using manual methods and tools, including laying out the type, placing amberlith for photo position, cropping photos, and drawing a ruled box; 2) compose another mechanical using PageMaker software according to specs; and 3) design an advertising or graphic product, including generating content ideas and executing them for several thumbnail sketches, several "roughs" (more developed than thumbnails), and one finished product, according to instructions provided. One recent year's creative project assigned a cover and first page for a children's cookbook, this year's was a table-top [tent-style] ad for a restaurant.

Carpentry

Contestants construct an element of a building, such as a stair stringer or a wall with a window, following blueprint instructions, for the main part of this contest; the other element is a written test that requires identifying tools, calculating relevant measures (e.g., cubic yards of concrete to fill a given space), and demonstrating general knowledge of accepted practices.

Practical Nursing

Tasks may include: obtaining/recording vital signs; changing a wound dressing; making an occupied bed; performing cardiopulmonary resuscitation and emergency cardiac care intervention; and preparing, administering, and recording medication following a doctor's instructions.

Job Skills Demonstrations (in Leadership Development contests)

Students can demonstrate any job skill that can be explained briefly; the goal is to teach the judges how to do the skill, so students must actually demonstrate it while describing or explaining how to do it. Components that are judged include organization of the presentation (including the presence of an appropriate introduction and conclusion); poise; clarity and grammar; diction, speed, control, and tone of voice; and overall content.

Figure E.1b—Some Specific Contests

otherwise subjective judgment (which comes up especially in cosmetology, advertising design, culinary arts, and even in skill areas like cabinetmaking and precision machining). In some cases, points are awarded for overall quality of a project, introducing a holistic element to the scores.

Competitions last over four days, including the opening and closing ceremonies (group competitions that test public speaking, organization, synchronized movement, and memorization). The competitors must wait in suspense to find out who won until the last night. At the awards ceremony, winners are announced in front of the thousands of conference participants, family members, and guests, and they go to the podium to receive their medals (gold, silver, or bronze. *All* contestants receive a certificate for having competed, and the other

top-ten finishers (those who place 4th through 10th in each contest) receive special recognition on their certificates.

The tests are all developed by technical committees. For the Leadership Development contests, these committees consist mainly of vocational-technical teachers, while in the Skills tests committee members are current or former professionals in the particular occupation or industry, along with some people who work for a relevant industry association or labor union. Despite the claims of the national office that none of the technical committee members in Skills areas are supposed to be vocational educators, there were some examples in 1995 of such members who were instructors or even administrators—not quite the same as working in the industry. For example, two members of the culinary arts committee were admissions directors, one was another type of administrator, and another was an instructor at a postsecondary school. It's highly likely that such committee members previously worked in the respective industry and occupation, but such practices do tend to undermine the claim that the contests are on the cutting edge of industry standards.

The technical committees develop the tests through a series of meetings and conference calls, usually also gaining feedback from teachers about whether planned aspects will work. The process is somewhat informal and varies considerably from one skill area to another. In only a very few areas, including electronics, precision machining, and technical drafting, the committees have used the standards produced by national skill standards development groups to guide their test content. The committees may draw on task lists for an occupation, or tests developed at the state or local level as well, though obviously they need to change the test questions substantially to prevent unfair advantage going to one or another state's competitors. (One technical committee member reported that a half-hour before the event began, he changed a factual element on the test [the type of metal, in automated manufacturing technology] that would change many of the calculations students needed to do, in order to guard against possible cheating.)

There was widespread agreement among sources interviewed on this point: the contests contribute significantly to VICA's overall goal, which is to prepare qualified and highly motivated workers for occupations, mostly in the skilled trade and industrial sectors, some of whom go on to become managers and leaders in these industries. In preparing for the contests and working with teachers who have access to VICA' s other offerings, students should gain not only the practical knowledge demanded by their industry but generic workplace skills and qualities needed for entry to and success in any industry (such as teamwork, dependability, integrity, decision making, and communication skills).

The specific role that the contests play, as opposed to the influence of VICA' s overall program, which includes teacher training workshops, publications, and access to personal networks in industry, will be discussed in more detail below.

Relationship to Other Programs

VICA aims for its contests to be closely related to instruction in classes taught by VICA members, though these ties undoubtedly vary across competition fields and across instructors. The tests seem to drive the curriculum content more than they respond to it. The skills and knowledge tested by VICA's Skill Olympics should in most cases also have been tested earlier, in the classroom, though the national contests may serve as a kind of final or comprehensive exam, not just for one course but for a program or group of related courses. VICA's contests bring together more content elements than a typical school test does, and may use more current or cutting-edge technology than some schools have (this can be a problem because some students will be less prepared than others through no fault of their own). VICA distributes the topics for next year's test in advance to teachers and student members so they have a good idea what to emphasize. Moreover, the specific competencies that will be tested are listed in the official regulations guide, which is revamped every three years.

In most of the skill areas, update seminars combined with training sessions are offered to teacher members through the state VICA offices on a regular basis, serving at least two purposes: first, keeping teachers informed about changes in the industry or occupation and any new curricular or teaching materials available; and second, gathering data from teachers about what topics they need more information on. At these sessions, teachers can discover tasks good for their own classroom tests and discuss curriculum and instructional issues with others from around the state (or country since seminars are also given at the national conferences). They also have opportunities to make contacts with industry representatives for field trips, internships, and mentorship opportunities. Teachers may even be able to earn college or inservice training credits by taking a lead role in running the state VICA competition or teaching continuing education seminars.

The links between classroom/workshop instruction and these national contests mean that students at the national competitions should be familiar with the format of the tests *and* should know in advance what content areas will be tested. For one thing, those who make it to the nationals have usually won in the same

area earlier in the school year at their state competitions.[1] In addition, teachers are supposed to have the official regulations book, which provides specific information about the competencies that the national contests may test. However, in practice many teachers do not have the current book; this problem was especially common among teachers who taught contestants in the Leadership Development areas. This lack of initiative on the part of some teachers is mystifying, given that the book costs only $10 and is updated only once in three years.

Implementation and Administration

The VICA contests started in the mid-1960s, with only three contest areas (all in Leadership Development), and have grown steadily since then. The national Skills and Leadership events now involve thousands of students in dozens of competitions. These assessments are implemented annually from the local to international level (only certain areas are competed in internationally, depending on the presence of strong programs in enough participating countries. Skill areas differ among the other levels of competition as well). VICA members in the state hosting the national conference, especially the teachers, play a stronger role than those from other states.

The specific content of tests is changed every year, though most of the subject material covered remains consistent from one year to the next. Students can compete in events for more than one year so long as they still meet the eligibility requirements, so VICA has to change the tests to keep things fair. Students must be active VICA members to compete and must compete during the school year in which they are enrolled in a course or program in the same subject area as the competition (for local, regional, and state competitions) or at the end of that year, in June (for the national competitions). Tests used at the national level are released for use by state, regional, and local competitions in subsequent years; these are usually modified and shortened because the state and local contests generally are finished within one day, so tests have to be shorter. Teachers can obtain these tests for use in their classrooms, too.

[1] The official rules state that only first-place winners from the state competitions go on to compete in the nationals, with the exception that if that person is unable to attend, the next-highest-scoring competitor takes his or her place. Moreover, discussions with teachers/advisors indicate that a similar sifting process is supposed to occur from local/regional competitions to the state level. However, in practice a small proportion of competitors have not had adequate preparation for the national contests, whether because of weak competition in their state or because they did not win in their state but were asked to represent the state anyway due to unusual circumstances.

All judges for the Skills tests are professionals in that field (or were recently): people who perform or supervise the tasks being tested or who set standards in the occupational field. Judges in the Leadership contests are teachers or professionals in the relevant field. All judges receive training in what to look for, how to score, how to strategize in assigning scores so that there's room above a strong performer for an even better or perfect score, and so on. To encourage consistency in judging, a rookie judge is always put on a team with experienced judges and encouraged to ask questions; the new judge's work is carefully monitored by the experienced judges. All judges attend a "familiarization session" for contests that have new equipment, computers, software, or a dramatically different test element; these sessions can last up to a full day. The extensiveness of the judges' training varies from one field to another, but no one interviewed suggested that the judges received less than adequate training.

People interviewed thought that each judge was fairly consistent across the projects she judged, even though there may be a wide variation across judges' scores for one contestant or on one test component. Since the scores used for ranking contestants are the totals from all judges, if one judge is consistently lenient this should not affect the relative position of any contestant. In some contests, one highly experienced judge will score all contestants' products or papers for a component of the test. Moreover, judging that may be subjective occurs on the job as well, and one line of thinking is that VICA student members should get used to that. National VICA staff say that there has been general acclaim for the fairness of the judging—but there are no hard data such as external reviews from unbiased experts to confirm this.

Technical Quality of the Assessments and Judging

No research has been done on the validity, reliability, or equity of VICA's assessments (either test content or methods). VICA has kept electronic records of the score results from the national contests for the last 5–6 years, so some of the data needed to do such research exists. However, there has not been sufficient demand for it to be done.

Validity. Those involved with VICA believe that the tests are highly valid because industry is extensively involved in designing them. Industry representatives work hard at the conference performing many different tasks (and on activities during the year leading up to it); they also secure funds and equipment because they have an incentive to train future workers who may contribute to their company's success, so they want to make sure that the tests are fair but challenging, up-to-date, and authentic to the tasks of that job.

Teachers also work hard on these activities, because they want to improve their students' chances for success, so everyone involved has strong, consistent motivations.

There are competitions in which logistical or other difficulties tend to obstruct the fairness of test administration, however. These are probably fairly rare occurrences that usually happen when a skill area is completely new or a new element is introduced to an existing contest. For example, in this year's advertising design contest, there were three problems that led to unfair advantage for certain contestants, in the view of one teacher/advisor. First, the technical committee did not allow adequate time or opportunity for competitors to clarify terms they hadn't heard before (they may have been familiar with the concept but by another name); allowing such discussion adds authenticity to the experience, since an ad agency employee would normally be able to clarify what the client wants before proceeding.

Second, part of the competition was done on computers at a local school, raising a couple of issues. Perhaps most significant, students who were familiar with the software used had a distinct advantage, especially since it was a timed project. (One could argue that if schools want to prepare their students for work they should have up-to-date computers and use the software most commonly used in the field, but obviously it is a matter of chance whether a given student had worked on that software. This chance occurrence reflects not on the student's abilities but on the school's resources.) Moreover, some of the printers failed to work properly, wasting time for some students but not others. In addition, a class was being conducted in another part of the room where some contestants were working, causing distraction—but for only some students.

Reliability: Informants thought scores were highly reliable because in many cases decisions about correctness of responses or performances are objective. In other cases, where subjective judgment enters into the score (particularly in aesthetic areas like how pleasing or effective a graphic design or advertising concept is, or how food looks and tastes, but also in how well the parts of a cabinet fit together or how close a written explanation is to the perfect answer), there is probably more variation among the scores from different judges. However, in the totals these differences likely even out (assuming that judges are consistent across the different contestants).

People running the contests strive for fairness. On one troubleshooting component of the electronic product servicing event, several contestants had a faulty reading on a diagnostic tool and thought this was the problem they were asked to identify. Although the person who designed this part of the test

thought they should have known this was not the intended problem, and said so at the debriefing, the technical committee reconsidered and eliminated that part of the test because several contestants had been misled in the same way and had filed a grievance to challenge their scores.

Debriefings are held in many contest areas to go over the correct answers, section by section, and discuss how contestants performed in the aggregate. At these sessions, students can learn from what they did wrong (and, to some degree, satisfy their curiosity about what their chances are of winning). Some of the debriefings stuck to the points raised in the contests, while others provided long-winded presentations of new materials and equipment, amounting to "free" endorsements/advertising for a company's products (not really free, since in exchange these companies donated items and labor for the conference). This is an unfortunate byproduct of the contests' corporate sponsorship, but it is minor compared with what the contests offer in opportunities to excel, to learn from others, and to discover more about employment and postsecondary training options.

Equity: Everyone asked about this issue said that it was not possible for a person's gender or race–ethnicity to influence judges of the Skills tests, since in many cases they fill out the scoring sheets for the whole group of products or answer sheets after contestants have departed; these are identified only by contestant number. However, in most of the Skills areas, the vast majority of the contestants were either white males or white females. In such a setting, the one or two contestants who don't fit the pattern (e.g., the one female welder among a group of males) may in fact stick in the judges' minds, since judges usually observe the performances as well. Thus, there may be subconscious bias either favoring or disfavoring such students who stand out from the group. In some trade areas, the groups were more mixed, especially by gender (e.g., advertising design and culinary arts).

In the Leadership Skills tests, judges observe the contestants one by one as they perform (except for those few tests that involve teams like parliamentary/chapter business procedure), so one can't argue that the procedures preclude bias. It is therefore possible for unfair judging based on gender or race–ethnicity to occur, though no one asked about this was aware of any such instance (it is also quite possible that judges take into consideration other factors beyond the skills and attributes tested, such as appearance or voice characteristics, whether consciously or not).

Consequences and Use of Assessment Results

Results are mainly informal, though this differs somewhat by the groups receiving or using those results. The discussion below addresses these groups separately.

Students: Though the goal is to share students' results with them, in practice they may have graduated from high school or their postsecondary program just before competing in June and may not be in touch with their teachers during the summer. Thus, many of the students don't find out their scores; in most cases, it's probably up to them to contact the teacher and ask for the score. The actual results are not used for other formal decisions about a competitor's schooling or employment; however, individual student competitors use the experience of preparing and participating in various ways: to make contacts for jobs or further training, to help them decide on particular avenues to pursue within an industry or job category (e.g., whether to strive to be a charge nurse, who manages other nurses in a unit), to bolster their resumes, to learn to handle pressure, and to learn skills or helpful tricks from other competitors' performances. (Students are more likely to have time to observe others in the Leadership events, where they can observe each competitor who follows their own performance, and in Skills tests that take relatively little time to complete and have more than one batch of competitors. In Skills tests that take most of the day, students have little or no time for observation.)

Short-term consequences for student participants are mainly informal effects on personality or character such as increased enthusiasm for schoolwork, higher self-confidence and aspirations, more assertiveness, gains in teamwork skills (for some events), and better decision-making ability. Results mentioned by students were generally quite positive, although students who had lost at the local, regional, or state level would probably paint a different picture. On the down side, some students experience anxiety or disappointment if they do not win, and a small minority even show temporary symptoms of illness from the stress. A small percentage of students are unpleasantly surprised to discover that industry standards represented at the VICA conference are higher (or cover different content) than standards at their school or at the state competition. A small proportion of students interviewed mentioned that they found it difficult to deal with strains on friendships when a good friend did not win a contest while they did, especially when they competed directly against each other (this is far more likely at state and local contests). These negative effects are minimal, however, compared with the substantial positive effects cited by all students interviewed.

Students from different backgrounds benefit from participating in the contests, though they may gain different things. An outstanding high school student in a college-prep or other highly rigorous program may sharpen her interpersonal, communication, and public-speaking skills through contest preparation or other chapter activities, while an average student in a general or vocational track may get involved in electronics or machining through VICA, do well and gain confidence, take more math classes, and decide to enroll in a technically oriented college program. There are even limited roles for developmentally disabled or other special needs students: the custodial services competition is restricted to students with individual educational plans. Several student competitors at the 1995 competition had a physical disability. Teachers thought that most students possessed a high degree of motivation before deciding to participate, while in a few cases the rewards offered by the VICA program may reach out to at-risk students. (One teacher of law enforcement/criminal justice issues from a poor, small-town school in Texas explained how he had used the lure of a steady job and higher chances of entering postsecondary school to turn around some potential gang members.)

Teachers and programs, others: Teachers are supposed to receive the individual scores for the student(s) competing from their program (one per skill area); the scores for each state's national competitors ranked, for each test component, plus the total scores; along with the national average and standard deviation of this distribution. Some teachers reported that they received information from their state director in a form that was difficult to understand, or had to track down the director to get the information. Others never receive the scores. (What is reported to teachers after the state and local competitions undoubtedly varies too.) Teachers who do receive readable results can see which skills they are teaching well or not so well, depending on how their student compares nationally.

However, to make a sound judgment about how they may need to change their curriculum, teaching methods, or equipment, a teacher really needs additional information; for example, a description or example of the ideal performance that the judges were looking for, whether their student was particularly nervous and did not perform up to potential, or whether another instructional factor harmed a student's performance (e.g., the school lacks a particular software program or type of equipment). At least one skill area, Precision Machining Technology, provides a detailed report that explains each item or element at each work station, including the perfect answer or product the judges were looking for and an overview of how students as a group performed in each station's activities (reporting subscore averages).

Teachers who observe the Job Skills contests may begin thinking about new components to emphasize in their teaching if they find that industry demands skills they weren't previously teaching or emphasizing sufficiently. Teachers can benefit from this knowledge whether or not they attend the national conference, either by discussing test content with participating colleagues or contestants who report back from the conference, or by obtaining tests used in previous years. Teachers are likely to revise their definition of acceptable performance (generally by raising their standards) after experiencing the competition at VICA. They are somewhat less likely to change teaching methods since instruction is not discussed or modeled at the conference (though they may learn new techniques informally by discussing them with other teachers).

"Teaching to the test" was a common effect of being involved with VICA, but since the tests are thought to reflect the knowledge and skills that industry wants in their staff members, this was seen as beneficial. (A side comment: some teachers, especially for some of the Leadership contests, apparently did not read the specific contest regulations carefully, or at all, and so were not in a position to teach their students to prepare well for the tests.)

Participating in other aspects of VICA's program: Teachers who join VICA have access to a range of benefits by attending seminars with industry representatives and statewide meetings, including:

- Curricular materials and training that focus on developing leadership and communication skills among the students, encouraging community service work, and improving attitudes toward school and work. In certain industries, national associations have produced new curricular materials (e.g. *Raising the Standard,* in electronic products servicing, developed by the Electronics Industry Association).
- Access to industry contacts and industry-led seminars and meetings (usually coordinated by the state director), where they can hear about new equipment and techniques, skills being sought by industry, and even labor market information.
- Continuing education credits, for those teachers who take a leadership role in helping to write or pretest an assessment, or soliciting support from industry, or planning for or managing a local or state conference.

Employers: Some employer representatives attend VICA partly to recruit employees and may talk to students they have seen performing especially well. Employers do not receive actual student scores from VICA, but they have access to the list of winners.

Applicability to Vocational Education

The costs of putting together a national contest like VICA's are enormous, covering three main areas: equipment, materials, and facilities; labor for contest design, setup, judging, and breakdown; and personal transportation to the convention site. Most of these costs would not be incurred if a school or district were to implement a similar assessment locally. The equipment and machinery required would presumably be available at the school; however, school-owned equipment would probably not be as up-to-date and possibly would be less sophisticated or useful than what is available at the national conference (these are all donated or loaned to VICA for the event). Additional materials beyond what a school's budget covers would likely need to be located for a competition (or in-class testing), either by purchase or by soliciting a donation from industry. The substantial facilities costs, which are paid for by VICA for the national competition (convention center rental, provision of plumbing connections and special ventilation ducts, electricity use, catering for certain events, etc.), could be avoided at local competitions. Personal transportation would similarly be mainly avoided (these funds are paid for by a variety of sources: students, their families, school fund-raisers, supporting firms, and in some cases school/district funds, especially for teachers' travel).

The most substantial obstacle in replicating these assessments in schools is replacing the labor that goes into it, especially test design and judging. Although this time is all contributed by professionals in the fields being tested and teachers, so there isn't a dollar amount that needs to be covered, it is unlikely that a local contest could begin to replicate this level of expertise. A key aspect to the value of VICA job skills contests is that the judging is done by expert practitioners, which gives both students and teachers a more realistic and current view of performance expected in the industry (compared with having teachers as judges). Moreover, the level of competition is bound to be higher with a larger pool of contestants, so students may not work as hard at preparation (as they do for national competition) if they know who their competition is (which they would if it's restricted to their class, or may if it's schoolwide) or if they have a sense that the competition is not very fierce.

However, despite these obstacles, the content of the tests can be useful to individual teachers and groups of teachers meeting to revise their courses or programs. The judging sheets are available in the USSO rules book (revised once every three years), and tests given at the nationals are released for use at state and district level competitions in subsequent years. The input from industry should also provide an important lesson for all vocational educators: even if it is

not possible to command the high level of commitment from industry representatives that VICA does in designing and implementing these assessments, schools should seek and use industry's viewpoints and knowledge wherever possible.

F. The Career-Technical Assessment Program

The Career–Technical Assessment Program (C-TAP) is being developed by the Far West Laboratory for Educational Research and Development for use in vocational programs in California's high schools and regional occupational centers/programs. (The California Department of Education is funding this program's development.) C-TAP assessments are being developed in five career areas: agriculture, business, health careers, home economics, and industry and technology education. Within each area, C-TAP is targeted either to clusters of occupations or the core course a student takes as an introduction to a specific career area, which teaches basic information relevant to a wider grouping of occupations. Today there are two of these C-TAP clusters or core courses for each career area:

1. Agriculture: core course, animal science
2. Business: marketing, computer science and information services
3. Health: introductory core course, advanced core course
4. Home Economics: child development and education, food services and hospitality
5. Industrial and Technology Education: core course, construction

Description and Purpose

Originally, in 1990, C-TAP was planned as a set of specific occupational tests for 29 occupations. The tests were to be made up primarily of multiple-choice questions and some performance items measuring specific skills for entry-level jobs. C-TAP's primary purpose was to be a standardized statewide student certification system; its secondary use was as a program evaluation tool.

C-TAP's purpose and content has changed over time. Early on, the focus switched away from specific occupations and job skills, and toward clusters of related occupations and broader skills. It was never used as either a student certification system or a program evaluation tool. Instead, C-TAP is currently being used both as a teaching/learning tool and as an assessment that contributes to the class grade in vocational education classes. It is not used in

any standardized fashion; rather, teachers adapt the materials for use in their classrooms. Far West Laboratory estimates that several hundred teachers are using C-TAP or the related generic skills version (CAP) this year.

Today, C-TAP has three components, each of which addresses academic skills, general workplace skills, and job-specific skills: 1) the portfolio, 2) the project, and 3) the scenario. Both the portfolio and the project have subcomponents. The portfolios and projects are to be graded by the teacher. Scoring is done holistically, with three possible grades: 1) Basic (unsatisfactory), 2) Proficient (very good), and 3) Advanced (excellent), using rubrics developed by Far West Laboratory. The scenario assignments will be distributed and scenarios scored by Far West Laboratory.

The *portfolio* includes a collection of materials that exhibit the student's skills, while producing the portfolio serves as a vehicle to motivate and help the student learn and polish skills (see Figure F.1). The portfolio must contain material in each of five parts. The first section presents the portfolio to readers, with a table of contents and letter of introduction for the student and his work. A career development section follows, including an application for employment or college, a letter of recommendation, and a resume. The third section contains four work samples, in which the student both illustrates mastery of specific skills relevant to the career area and writes about her understanding of the skills' importance. The fourth section is a writing sample related to the student's career area. Fifth is an evaluation of supervised practical experience, which may or may not be required by specific career programs but is encouraged for all students.

The *project* component is intended to bring together a large body of the student's work, including hands-on activities done by students individually or in small groups, in order to show the development of specific career skills. The project is to be done in four stages, each of which is evaluated. First is the project's plan, and second is evidence of progress towards completing it. Third comes the completed project itself. Fourth is an oral presentation on the project. Originally, C-TAP did not include a project but instead required an on-demand performance task (chosen centrally for all schools) and a separate oral presentation. When pilot-tested, the on-demand task met with opposition from teachers who felt it did not reflect instructional and curricular differences among schools. So the task and the oral presentation were merged together to create projects, which allow some student choice, and which teachers saw as more likely to fit with their curriculum.

The *scenario* is a 45-minute essay test in which the student is presented with a real-life problem in their career area and is required to evaluate the problem and

The work samples in the portfolio must document one or several specific technical skills and show the student's ability to communicate information about the skill. The samples combine written material plus illustrations (drawings, or photographs of physical artifacts the student created). Students also write a description of the work sample and explain the skills it demonstrates.

A student in a child care and development class designed a day-care center as one of his work samples. The design included a floor plan of the center, a list of criteria to evaluate safety problems, a list of items to stock the center (including those needed to meet children's developmental needs), plus each item's cost.

A student in an animal science class documented how she gave a cow with mastitis an intramammary injection, using drawings, a picture of herself with the cow, and a written description of the procedure.

In a class for dental assistants, work samples have included documenting how a student sterilizes instruments, takes a dental impression, pours a plaster mold, and takes a full set of x-rays.

Figure F.1—Examples of Work Samples from Portfolios

A student developed a travel brochure for Mexico as his project for a business class. First, he laid out a 12-step plan to develop the brochure and identified the resources necessary to implement the plan. As evidence of progress, he submitted a journal of his activities for developing the brochure, an interview protocol he used with travel agents, a data base of vacation hotels and their prices, and information on vacation spots located on the Internet. His final product was an eight-page brochure describing three vacation spots in Mexico (Los Cabos, Mazatlan, and Puerto Vallarta), bringing together the information he had supplied as evidence of progress.

A student in a core agricultural science class reforested an area of marginal grazing land. First, he laid out a five-step plan and identified necessary resources. As evidence of progress, he submitted a journal that showed activities over four months, including land preparation, purchase of 200 trees, planting, and a site visit three months later to examine the trees. His final product was a photographic journal of the site, its preparation, and tree planting.

A student in a health class displayed the method of autopsy most commonly used by medical examiners. She set out a 14-step plan and list of resources. Her evidence of progress included a journal of activities, photos of herself and a partner drawing a model of a human torso, photos from a trip to a medical examiner's office, and an outline of information on autopsies. Her final product was three photos of a cadaver in various stages of an autopsy, with captions describing the stages, and a three-page written description of autopsy procedures.

Figure F.2—Examples of the Project

propose a means of addressing it. Scenarios were included in the C-TAP to provide an on-demand task and to directly address students' problem-solving skills. A veterinary science class provides one example of a scenario. Students read a description of an unhealthy cow's symptoms and the conditions in which it was kept. They were asked to identify the illnesses the cow may be suffering from, their causes, and possible treatments for the cow's ill health.

Relationship to Other Programs

Along with the purposes above, in the future C-TAP may return to its original intended use, as part of a certification program. The state has already incorporated it into certain reform initiatives. For example, the state requires that 80 sites having tech-prep programs use C-TAP in assessing student progress. So far, the state department of education has given little guidance on this required use of C-TAP; in contrast, schools and programs score C-TAP themselves and can use it to develop their own certification system. Additionally, the state is considering the use of Certificates of Initial and Advanced Mastery; if this initiative is adopted, C-TAP may become a requirement for the receipt of one or both certificates.

The C-TAP will be linked to two sets of state standards: 1) career-technical model curriculum (content) standards, and 2) career preparation (generic workplace readiness) standards. Both sets of standards were developed by the California Department of Education with technical assistance from Far West Laboratory. The career-technical model curriculum standards are in the process of being formally adopted by the state Department of Education (see the C-TAP Cluster-Specific Supplements 1995). There are separate standards for each occupational cluster. Both the work samples in the portfolios and the scenarios are linked directly to these standards. The writing samples and projects are less closely linked to them. The seven career preparation standards apply commonly to all five career areas (see Appendix B of the C-TAP Teacher Guidebook 1994). The supervised practical experience evaluation in the portfolio asks for the student to be evaluated on these standards.

Implementation and Administration

All of the teachers interviewed used the portfolio component, and it was generally deemed the strongest part of C-TAP. In turn, teachers saw the work samples as the most valuable part of the portfolio. Work samples require the student to explain what they have learned, which reinforces the material while also giving the student a record of what she has learned for the future. This technique both bolsters the student's self-esteem and can be shown to others as evidence of the student's skills.

Teachers believe that three types of skills are required for the portfolio: applied academic skills (especially writing), work readiness skills (demonstrated in the supervised practical experience and in how well the student communicates in the work samples), and specific technical skills. The requirements for the portfolio

vary by class. For example, the number of work samples included may be more than the recommended four. The level of detail in work samples can vary greatly, e.g. from a half-page description of how to use a fax machine to several pages with photos describing how to vaccinate a sheep. The requirement for the writing sample also varies and has included a research paper written for the class, a paper related to the subject written in another class, and an expository paper describing a personal experience.

Teachers also differed in the weight they placed on the C-TAP for evaluating students. All the surveyed teachers graded the subcomponents of the portfolio (with primary emphasis on the work samples and writing sample) and the portfolio overall. The contribution of the portfolio to the student's overall grade varied by teacher but reached as high as 80 percent. The majority also required the submission of a completed portfolio in order to pass the course, and one teacher tied a completed portfolio to receiving the career certificate plus college credit for the high school course. Where used, the projects were graded. Scenarios might be graded or not, depending on the teacher.

The scenario and project components of the C-TAP are being used less consistently than the portfolio. Reasons teachers gave for not using scenarios or projects include: they create too much work when combined with the portfolio; they're already assigning and evaluating projects, so the C-TAP project materials don't add anything new (or not enough that's new); or the time gaps between teaching particular material and receiving the relevant scenarios from Far West. When teachers use projects, they often simply use the ideas provided in C-TAP to give more structure to an existing assignment. Use of the scenarios varies from using them once or twice a class to using them more often for practice, or once at the end of a unit.

Technical Quality

Far West Laboratory's first step in determining the reliability of C-TAP has focused on developing the means to consistently score the assessments. So far, this work has entailed creating benchmarks, primarily for portfolios but also some work for projects. (Benchmarks are examples of a student product that merit each of the three ratings: Basic, Proficient, and Advanced. They are used to train teachers how to rate the products.)

In the summer of 1994, teams of 8–10 teachers spent two days identifying benchmarks for portfolios and projects, together covering all five career areas. All members of the group had received portfolios for review before coming

together. In the teams, they discussed the pieces of work before them and agreed on which portfolio dimensions were important to scoring. They were then trained in applying Far West Laboratory's scoring rubric with its three scores of Basic, Proficient, and Advanced. Their job was to rate additional portfolios to find good examples of each score. To help them in this work, teachers were allowed to rate the portfolios using five scores (Basic, Basic-Proficient, Proficient, Proficient-Advanced, and Advanced), but only examples of the three desired scores were to be selected and presented to their team. The team agreed upon two examples to be the benchmarks for each rating level (except that in some cases they couldn't find two examples for the Advanced rating). An additional day was spent determining that the benchmarks were equivalent across program areas. In all, the teams developed benchmarks for portfolios for one occupational cluster (or the basic core class) in each of the five program areas:

1. Agriculture: animal science
2. Business: computer science and informational services
3. Home Economics: child development and education
4. Health Careers: core class
5. Industry and Technology Education: construction

Implementation of C-TAP started slowly during school year 1994–95: portfolios were begun only in January and were due in May. That summer (1995), when teachers convened to select additional benchmarks, many of the portfolios turned out to be incomplete or problematic; their contents had to be cleaned up or combined to make a full portfolio. Far West Laboratory calls these examples, which teachers helped produce, "exemplars" rather than benchmarks. Exemplars were developed for projects and portfolios, but work is not yet finished using the 1994–95 materials. Exemplars were developed for 10 occupational clusters, two in each of the five program areas:

1. Agriculture: animal science; basic core class
2. Business: computer science; marketing
3. Home Economics: child development and education; food science and hospitality
4. Health Careers: two core classes
5. Industry and Technology Education: construction; core class

By December 1995, Far West Laboratory had drafted a set of examples of student work considered proficient for both portfolios and projects for the 10 occupational clusters. The examples were to be used by teachers and students.

For the summer of 1996, Far West Laboratory plans to repeat the process of developing benchmarks using whole portfolios and projects. The benchmarks will be used by teachers rating projects and portfolios; these ratings will in turn be used to determine interrater reliability and internal consistency of student scores. Additionally, there are plans to score the same materials again, using an analytic rubric, for three purposes: checking interrater reliability for this rubric, measuring the usefulness of subscores provided by this rubric, and comparing the reliabilities of the holistic and analytic scoring rubrics.

Scenarios have been examined using a different process. In the summer of 1995, teachers came to Far West Laboratory and rated scenarios to determine how well the scenarios worked. They found that students gave such different answers that scoring was difficult to do. Far West Laboratory is thus revising the scenarios and plans to have students take them during the 1995–96 school year, after which they will do a large-scale rating of scenarios to determine rating reliability. Currently, Far West staff are determining whether the rubric for scenario scoring should remain holistic or change to an item-specific approach. Further research is planned to examine the number of scenarios necessary to obtain acceptable reliability for an individual student's score.

To address content validity, Far West Laboratory originally assigned two committees of experts the task of determining the type of performance assessments to include. These committees were influential in the decision to drop multiple-choice questions. There was a steering committee made up of the state vocational education director and the program managers for the five career areas. Also, there were advisory committees for the five career areas (some areas had more than one committee) composed of staff from the state department of education, career–technical teachers, academic teachers especially in English and the language arts, and employers.

The advisory committees were replaced by development committees (the memberships overlapped), each focusing on a specific occupational cluster. There are 10 development committees today, made up primarily of career-technical teachers but also including academic subject teachers, postsecondary faculty, and employers. Focusing mainly on portfolios, these committees have produced lists of acceptable topics for portfolios, described the structure of the work samples, and outlined what supervised practical experiences should include. Far West Laboratory developed the assignments for portfolios using input from these committees while ensuring that they would be equivalent across occupational clusters and career areas.

A further step toward checking content validity: focus groups have examined completed student work to determine whether it demonstrates the skills desired in the workplace. Far West Laboratory has received positive reports from these groups. Because this work has been somewhat informal, the Laboratory plans to establish more formal focus groups made up of employers, industry practitioners, teachers, and postsecondary educators in the relevant career field. These focus groups will consider whether the current form of each component of the C-TAP and its scoring process meet the model originally set out. No other validity work has been done, such as correlating C-TAP ratings with other student ratings (such as test scores or teacher ratings) or using C-TAP ratings to predict student performance on particular criteria. But Far West Laboratory does have plans to correlate C-TAP ratings from the 1995–96 materials with student scores on course tests, course grades, GPA's, and scores on standardized tests. Additionally, there are plans to examine differences between students who completed the C-TAP and those who did not, for example, in acceptance and enrollment rates in postsecondary schools or other training.

Uses of the Results

Since the C-TAP was never used as a student certification system, it has not been implemented in a standard manner at all schools. Teachers thus differ in their use of the three C-TAP components and in the importance they place on the C-TAP work in evaluating students.

The C-TAP, especially the portfolio, is intended to record a student's learned skills. The portfolio has been promoted as a means of impressing employers and postsecondary institutions with the student's skills. Teachers had a somewhat skeptical view of the usefulness of the C-TAP for these ends. At present, they believe that postsecondary institutions are generally unwilling to use portfolios for making admissions decisions. One teacher did note that colleges were more willing to give her high school course college credit upon reviewing her students' use of portfolios. Views of the assessments' usefulness for job-seeking were more mixed. Anecdotal evidence suggests that some employers are impressed by portfolios but others are not. Overall, they thought it was a challenge to get employers to consider portfolios when hiring. Teachers argue that having a portfolio gives a student examples of their work that they can discuss with an employer and show if there is interest. Where students do internships while taking classes, C-TAP may also generate employer interest. For example, a health teacher noted that several dentists had asked for student interns who would be producing work samples as part of their internship.

To date, the purpose of the C-TAP is still evolving from the original idea of a student certification program. To different degrees in different schools, it has modified instruction, become a source of information for grading or a requirement for passing a class, and created a record of student work that may convince both a student or a prospective employer that the student has learned useful skills. Far West Laboratory plans to evaluate C-TAP's impacts on instruction and curriculum by surveying teachers, reading their journal entries, and reviewing teaching materials.

Applicability to Vocational Education

The feasibility of implementing C-TAP is difficult to gauge for two reasons. First, C-TAP has not been adopted on a schoolwide basis yet. Therefore this section focuses on individual teachers' implementation work. Second, not all of the teachers surveyed have adopted all three components of C-TAP. Some argue that there is not enough time to do all three and others argue that they will adopt the project and scenario only after they have become comfortable with the portfolio. There is also some concern that scenarios would be used to evaluate teachers, as they are centrally scored. For these reasons, this section focuses on the feasibility of implementing C-TAP portfolios. Feasibility here includes the effects of portfolio use on teacher time and responsibilities, plus the reaction of teachers, parents, students, and schools.

Teachers agreed that using the portfolio took substantial amounts of class time, especially in the first year. Much of the time went to explaining what was expected and giving students adequate time to carry out certain parts of the portfolio that traditionally were not part of the class (e.g., resumes and letters of application). For most teachers, the work samples came from existing activities but still required extra time for the writing of summaries. Teachers did give up teaching some material to make time available for the portfolios.

Also, additional, or at least different, teaching demands are placed on teachers who use the portfolios. They have to focus more on writing, especially specialized writing like résumés and job/college applications; teachers must also write recommendation letters for many or all students in a class. Additionally, class management skills are important, because students complete work samples at different speeds. Grading does not appear to pose an additional burden when the portfolio is used to supplant traditional tests or practicals. In fact, teachers said grading portfolios was easier because it's obvious from the work samples whether the student understands the material. If the traditional assessments are maintained as well, however, C-TAP would require additional grading time.

A further demand placed upon teachers is creating storage space for the portfolios. Physically, portfolios take up a lot of classroom space and students need to have easy access to them. Some teachers want to keep all work samples done in class, not only those picked for the portfolio, so that students can change their portfolio's content. Other teachers are trying to determine where to maintain portfolios that will be kept over all the years students are in the program.

Adoption of the portfolio has only occurred on the individual teacher level so far. The teachers surveyed believe the portfolio is a valuable approach and see a reduced learning curve for other teachers who subsequently adopt it with some help from them. Some have been involved in training large numbers of their colleagues in its use and believe that teachers with different levels of experience will vary in their willingness to use it. New teachers may not have time to learn to use it, while older teachers, especially those near retirement, may not want to invest the time. They do not yet see strong support from school administrators for using the portfolios. At first, students seem to resist the additional work requirements of the portfolio but over time they come to accept them; some students value having the record of work when they approach employers. Teachers have kept the portfolios at a fairly low profile as they learn to use them. For this reason, there has been little community response to them. Parents are generally not familiar with them, but when teachers have explained them, parents usually react favorably.

In conclusion, widespread adoption of portfolios by individual teachers appears to be feasible, if teachers can find sufficient time to learn how to use them and agree that substantive course material may have to be dropped to free up time. Far West Laboratory is planning to increase training of trainers, both Department of Education personnel and teachers, to expand the number of teachers who can be trained to use C-TAP. Currently, Far West is distributing the Cluster-Specific Supplements (which contain the career–technical model curriculum standards, ideas for projects and writing samples, and an example of a work sample writeup), and the Examples of Proficient Student Work (each of which contains one portfolio or project; together these cover the occupational clusters). Teachers and students can use these to gain a better understanding of how to implement C-TAP and what completed projects and portfolios should look like.

References

Boesel, D., and McFarland, L. (1994, July). National Assessment of Vocational Education: Final report to Congress. Volume 1: Summary and recommendations. Washington, D.C.: U.S. Department of Education

Bishop, J. (1995).

Hambleton, R. K., Jaeger, R. M., Koretz, D., Linn, R. L., Millman, J., and Phillips, S. E. (1995). Review of the Measurement Quality of the Kentucky Instructional Results Information System, 1991–1994. Frankfort: Office of Educational Accountability, Kentucky General Assembly.

Herman, J. L., Aschbacher, P. R., and Winters, L. (1992). A practical guide to alternative assessment. Alexandria, VA: Association for Supervision and Curriculum Development.

Hill, Clifford and Larson, Eric, "Testing and Assessment in Secondary Education: A Critical Review of Emerging Practices," NCRVE, December 1992.

Hoover, H. D., and Bray, G. B. (1995). The research and development phase: Can a performance assessment be cost-effective? Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Kentucky Department of Education. (1993a). Kentucky Mathematics Portfolio: Teacher's Guide. Frankfort: Author.

Kentucky Department of Education. (1993b). KIRIS: 1991–92 Technical Report. Frankfort: Author.

Kentucky Department of Education. (1995a). KIRIS: 1992–93 Technical Report. Frankfort: Author.

Kentucky Department of Education. (1995b). KIRIS Accountability Cycle I Technical Manual. Frankfort: Author.

Kentucky Institute for Education Research. (1994). A Review of Research on the Kentucky Education Reform Act (KERA). Frankfort: Author.

Kentucky Institute for Education Research. (1995a). An Independent Evaluation of the Kentucky Instructional Results Information System (KIRIS). Frankfort: Author.

Kentucky Institute for Education Research. (1995b). Summary of Research Related to KERA. Frankfort: Author.

Koretz, D. Linn, R. Dunbar, S. and Shepard, L. (1991). The effects of high stakes testing on achievement. Presentation at the annual meeting of the American Educational Research Association, Chicago.

Koretz, D., Stecher, B., Klein, S., McCaffrey, D. and Deibert, E. (1993, December). Can portfolios assess student performance and influence instruction? The 1991-92 Vermont experience. CSE Technical Report 371, Los Angeles, CA: CRESST/UCLA. (Reprinted as RAND, RP-259, 1994.)

Lazerson, M., and Grubb, W. N. (1974). American Education and Vocationalism: A Documentary History 1870–1970. New York: Teachers College Press.

Levesque, K., Premo, M., Vergun, R., Emanuel, D., Klein, S., Henke, R., Kagehire, S. and Houser, J. (1995). Vocational education in the United States: The early 1990s. NCES 95-024. Washington, D. C.: U.S. Department of Education.

Mehrens, William. (1992, Spring). Using Performance Assessment for Accountability Purposes. Educational Measurement: Issues and Practice 11(1):3-20.

Miller, M. and Legg, S., "Alternative Assessment in a High-Stakes Environment," Educational Measurement: Issues and Practice 12(2):9-15, Summer 1993.

MPR Associates. (1996). Skill Standards: Concepts and Practice in State and Local Education. Berkely, CA (unpublished).

Rahn, M. L., Alt, M., Emanuel, D., Ramer, C., Hoachlander, E. G., Holmes, P., Jackson, M., Klein, S. and Rossi, K. (1995, December). Getting to work. Module Four: Student Assessment. Berkeley, CA: MPR Associates.

Shavelson, R. J., Gao, X. and Baxter, G. P. (1993). Sampling variability of performance assessments. Journal of Educational Measurement, 30, 215-232.

Shepard, L. (1991). Will national tests improve student learning? Phi Delta Kappan, 71, 232-238.

Shepard, L. and Dougherty, K. (1991). Effects of high stakes testing on instruction. Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, New Orleans.

Smith, M. L. and Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. Educational Measurement: Issues and Practice, 10(4), 7-11.

Stecher, B. M. (1995). The cost of performance assessment in science. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

U. S. Congress, Office of Technology Assessment (1992). Testing in American schools: Asking the right questions. Author.

Vocational Education Journal, 199?

Wiggins, G. (1989, May). A true test: Toward more authentic and equitable assessment. Phi Delta Kappan, 70(9), 703-713.

Wolf, D. P. (1992, May). Good measures: Assessment as a tool for educational reform. Educational Leadership, 49(8), 8-13.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

☒ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").