DOCUMENT RESUME

ED 355 193	SP 034 279

AUTHOR	Stansbury, Kendyll; Long, Claudia
TITLE	Assessment Component of the California New Teacher Project: Summary Evaluations of Innovative Assessment Methods.
INSTITUTION	Far West Lab. for Educational Research and Development, San Francisco, Calif.
SPONS AGENCY	California Commission of Teacher Credentialing, Sacramento.; California State Dept. of Education, Sacramento.
PUB DATE	Feb 92
CONTRACT	TCC-CO21
NOTE	143p.; For related documents, see ED 323 197, ED 342 761, and SP 034 278-282.
PUB TYPE	Reports - Descriptive (141) -- Tests/Evaluation Instruments (160)
EDRS PRICE	MF01/PC06 Plus Postage.
DESCRIPTORS	*Beginning Teachers; *Cost Estimates; Elementary Secondary Education; *Evaluation Methods; Evaluation Research; *Evaluation Utilization; Higher Education; Measurement Objectives; *Measurement Techniques; Preservice Teacher Education; Teacher Certification; *Teacher Evaluation
IDENTIFIERS	*California New Teacher Project

ABSTRACT

This report summarizes conclusions from previous reports and makes explicit comparisons between assessment approaches. The report is organized in four sections under the headings: Design, Content Analysis, Costs, and Analysis of Technical Quality. The first section (Design) describes salient characteristics of each assessment approach, the instruments pilot tested, and the type of teaching skill evaluated by each instrument. Section 2 (Content Analysis) uses previous analyses and experiences to draw conclusions about the congruence of assessment instruments with State Policy documents and about the ability of each assessment approach to assess specific domains of a teacher's knowledge and skills. The third section (Costs) presents an overview of the administrative complexity and of the scoring complexity of each approach. Cost estimates for administration are provided and the degree of developmental work needed is discussed. The fourth section analyzes the technical quality of the assessment approaches chosen and the appropriateness of each approach for diverse potential uses (licensure/certification, hiring and retention, and professional development). The conclusions of the study indicate that there is no one best assessment approach. The choice of the optimal assessment approach depends on the skills to be assessed and the purpose for which the information is to be used. If a broad examination of teaching competence is desired, then a combination of approaches is necessary. Ten tables are included in the report. (Contains 25 references.)

This report is based upon work performed pursuant to Contract TCC-C021 with the California Interagency Task Force of the California Commission on Teacher Credentialing and the California State Department of Education New Teacher Project. The contents do not necessarily reflect the views or policies of these agencies and the accuracy of the information is the sole responsibility of the Far West Laboratory.

Assessment Component of the California New Teacher Project:

Summary Evaluations of Innovative Assessment Methods

Kendyll Stansbury
Claudia Long

February, 1992

Gary Estes, Project Director
Far West Laboratory for
Educational Research and Development
San Francisco, California

TABLE OF CONTENTS

TABLES

INTRODUCTION

This report summarizes the experience of the California New Teacher Project (CNTP) in exploring alternative approaches for the assessment of beginning teachers. The CNTP was created by the legislature in the Teacher Credentialing Law of 1988 (Chapter 1355 of the Statutes of 1988.) Charged with exploring innovative methods of beginning teacher support and assessment, the CNTP occurs in the context of several other reform efforts in K-12 education and teacher education in California, including:

- the development of Model Curriculum Guides and Frameworks for subjects taught in elementary and secondary schools which call for more complex instructional strategies emphasizing active learning and the teaching of higher-order thinking skills;

- revised expectations for the support and evaluation of prospective teachers, as reflected in a series of *Standards of Program Quality and Effectiveness* for teacher credential programs in colleges, universities, and school districts; and

- revised requirements for the demonstration of subject-matter knowledge by prospective California teachers as reflected in a new battery of subject-matter knowledge examinations that include performance assessments.

The CNTP, jointly administered by the California Department of Education and the Commission on Teacher Credentialing, has three components: Support, evaluation, and assessment. The support component of the CNTP consists of local pilot projects in diverse teaching contexts which utilize a variety of approaches to support beginning teachers, as well as different funding sources. While these projects are not the only beginning teacher support programs in California, teachers and administrators in these projects are participating in the research on alternative methods of beginning teacher support sponsored by the CNTP. The cost-effectiveness of the various methods of support, and their effects on beginning teacher effectiveness and retention have been investigated by the evaluation component of the CNTP. The Southwest Regional Laboratory (SWRL) was contracted to perform the evaluation activities to identify the forms of support and degree of intensity of assistance that are most effective with beginning teachers entering the profession (Dianda et al., 1990, 1991; Ward et al., 1992).

Assessment Component

Critics have argued that multiple-choice tests should not be the sole means of evaluating teaching (or student learning), as they cannot fully assess many important skills and abilities, particularly those related to active learning and higher-order thinking skills. They have recommended the exploration of other approaches for evaluating the capabilities of teachers which are more authentic with respect to teachers' actual duties and requirements. Approaches which are considered to be more authentic include on-site observations, oral interviews, structured exercises in assessment centers, and the use of portfolios, videotaped scenarios and other materials in performance assessments. When education policymakers in California faced the choice of assessment approaches, however, they quickly discovered that few, if any, of the recommended methods had been pilot tested or evaluated in practice. Therefore, the CNTP includes an assessment component that was designed to evaluate innovative forms of beginning teacher assessment.

The first year of the Assessment Component of the CNTP concentrated on identifying and pilot testing the most promising existing assessment instruments for teaching. Five instruments were identified, including a classroom observation system, three semi-structured interviews, and an innovative multiple-choice test. All but one of the instruments, a semi-structured instrument for master teachers in social studies that was deemed too difficult for beginning teachers, were subsequently pilot tested. Having exhausted the supply of existing instruments considered promising, a request for proposals to develop additional assessment approaches or to adapt existing instruments to add an additional emphasis (e.g., subject-specific pedagogy) was circulated. Five instruments were selected for development and subsequently pilot tested. These include one classroom observation instrument with an emphasis on subject-specific pedagogy, two structured simulation task instruments, an instrument for analyzing responses to videotaped teaching episodes, and a set of performance-based assessment center exercises, including a portfolio component. Overall, twelve instruments representing seven different assessment approaches were pilot tested.

The analyses of individual instruments are contained in two reports of the CNTP Assessment Component (Estes et al., 1990; Estes et al., 1992), which concentrated on analyzing each instrument separately. The present report builds on these individual analyses by comparing the assessment approaches represented by the instruments on a number of dimensions. These comparisons, however, do not yield a rank ordering of the assessment approaches on a continuum from best to worst. The specific domains of teaching to be measured and the purpose for which the assessment is to be used strongly affect the evaluation of each approach.

DESIGN

In this report, the comparison of assessment approaches is grounded in the pilot testing experience. Three of the seven assessment approaches are represented by more than one instrument. One of the remaining approaches is an innovative form of the multiple-choice examination approach, which has been studied extensively as an assessment approach. Having two to three instruments represent an assessment approach helped to distinguish strengths and weaknesses of the approach from those of the particular instrument, as well as to identify features which could be modified in the future to strengthen the approach. Although often in the initial stages of development, the instruments were considered, at the time of the pilot testing, to be state-of-the-art exemplars of their assessment approach. It is possible that unforeseen breakthroughs can improve an assessment approach to overcome some of the limitations identified in this report. For example, one assessment approach -- videotaped teaching episodes -- was represented by a prototype instrument that was designed to reduce costs. However, this instrument did not seem to fully capitalize on the assessment methodology. Therefore, conclusions about the videotaped teaching approach are considered tentative. Each of the other assessment approaches was represented by at least one instrument which seemed to exploit well the assessment methodology. For these approaches, major improvements would require some sort of methodological reconceptualization of either the stimulus materials or the scoring approach.

Overview of Assessment Approaches and Instruments Analyzed

Twelve instruments representing seven assessment approaches were analyzed. This section describes salient characteristics of each assessment approach, the instruments pilot tested, and the type of teaching skills evaluated by each instrument. In addition, the report which analyzed each instrument is identified. Those approaches which are classroom-based (i.e., high-inference classroom observations and portfolios) are discussed first. Performance simulation approaches (i.e., semi-structured interviews, structured simulation tasks, and performance-based assessment center tasks) are discussed next, followed by other assessment approaches (i.e., videotaped teaching episodes and multiple-choice examinations.)

Classroom-based Assessment Approaches

Classroom-based assessment approaches evaluate a teacher's skills in the context of a classroom of students for which the teacher has continuing responsibilities. This type of approach includes both high-inference classroom observations and portfolios.

High-inference classroom observations. A classroom observation approach to teacher assessment consists of observing teachers as they instruct students in their classrooms. High-inference instruments specify rating categories in general terms, e.g., "monitoring and adjusting instruction," and rely heavily on observer judgment in rating. By contrast, low inference instruments specify the categories narrowly, e.g., "praises student," with the observer either counting the frequency of behaviors or noting the presence or absence of each behavior specified. Two high-inference classroom observation instruments were pilot tested. One, the Connecticut Competency Instrument (CCI), assessed ten teaching competencies. All of the competencies focussed on general pedagogy with the exception of one which examined lesson content. The CCI was evaluated in the *First Year Report* (Estes et al., 1990).

A second instrument, the Science Laboratory Assessment, was modeled after the CCI, but included categories to examine the effectiveness of instruction in a specific subject area, laboratory science. Seven teaching competencies were assessed, three of which were in the area of general pedagogy. Content knowledge, subject-specific pedagogy, and knowledge of students were also assessed. The Science Laboratory Assessment was evaluated in the *Second Year Report* (Estes et al., 1992).

Portfolios. A portfolio is the documentation of actual teaching experience, either examples of what the teacher considers to be superior work (e.g., lesson plans, handouts, student work produced after instruction), or materials related to an actual unit taught. The pilot tested portfolio took the latter form. Teacher portfolios in the pilot tested instrument were assessed according to six categories: Planning, curriculum framework, presentation, general pedagogical abilities, subject-specific pedagogical abilities, and reflection. The emphasis of the portfolio was on subject-specific pedagogy, but some aspects of general pedagogy were addressed as well. Although the portfolio assessment accompanied the exercises in the performance-based assessment center exercises (described later), and was an integral part of the overall assessment, it is discussed separately, because it represents a distinct assessment approach. The portfolio portion of the Secondary English Assessment was analyzed in the *Second Year Report*.

Performance Simulation Assessments

Performance simulations consist of tasks or exercises which simulate the performance of actual teaching responsibilities. The teacher is required to produce a product, such as a lesson plar, or to make a decision, such as selecting an instructional approach to use in a particular situation or deciding how to solve an instructional problem. Performance simulation assessments evaluated in this project include semi-structured interviews, structured simulation tasks, and performance-based assessment center exercises.

Semi-structured interviews. One form of performance simulations are semi-structured interviews, which provide opportunities for teachers to respond orally to a standardized set of questions or tasks that are presented verbally by an interviewer who uses a script known as an interview schedule. Semi-structured interviews can include "probes" that are used at the administrator's discretion to ask candidates to elaborate on their responses. Three examples of semi-structured interviews were pilot tested: The Semi-Structured Interview in Elementary Mathematics (SSI-EM), the Semi-Structured Interview in Secondary Mathematics (SSI-SM), and the Semi-Structured Interview in Secondary Social Science (SSI-SSS.) The SSI-SSS was an attempt to replicate the methodology used by the SSI-SM in a different subject, so the two assessments are parallel in many respects. Two tasks were common to all three interviews: Lesson planning and structuring a unit (called topic sequencing in the SSI-EM.) The SSI-EM also included tasks which required responding to student questions or remediating student errors and analyzing proposed computational short cuts. Both the SSI-SM and the SSI-SSS included tasks on evaluating alternative pedagogical approaches to the subject and on evaluating student learning. In addition, the SSI-SM included a task on analysis of alternative strategies for solving a mathematical problem, and the SSI-SSS included a task on teaching students historical interpretation. The SSI-EM focussed mainly on subject-specific pedagogy with some emphasis on content knowledge, while the SSI-SM and the SSI-SSS examined content knowledge, subject-specific pedagogy, and knowledge of students. The SSI-EM and the SSI-SM were evaluated in the *First Year Report*, while the SSI-SSS was evaluated in the *Second Year Report*.

Structured simulation tasks. Structured simulation tasks require teachers to analyze a completed teaching task, to outline how they would perform a task, or to actually perform a task that simulates one or more teaching responsibilities. The teacher's response is then compared to a list of previously identified responses or response characteristics. Three instruments composed of structured simulation tasks were pilot tested: The Assessment of Competence in Monitoring Student Achievement in the Classroom, the Secondary

Life/General Science Teacher Assessment, and Structured Simulation Tasks for Secondary English Teachers. The latter two used the same methodology and were developed by the same assessment developer. Two forms of the Assessment of Competence in Monitoring Student Achievement in the Classroom, each consisting of ten exercises, were pilot tested. These assessment instruments focussed on diagnosing and evaluating student achievement in the elementary grades. Five tasks of the Secondary Life/General Science Teacher Assessment were pilot tested: Applying effective instructional techniques, teacher as curriculum decision-maker, parent/student letter, lesson planning, and classroom and facility safety. Five tasks were also pilot tested as Structured Simulation Tasks for Secondary English Teachers: Designing a lesson sequence, developing oral presentation skills, stages of the writing process, responding to student writing, and responding to typical problems. The main emphasis in each of these tasks was on subject-specific pedagogical skills, with some additional attention to general pedagogical principles and parent/teacher relations. The three examples of structured simulation tasks were analyzed in the *Second Year Report.*

Performance-based assessment center exercises. Performance-based assessment center exercises have two characteristics: (1) they bring teachers together at a central place to participate in a series of activities, each of which uses a different methodology to measure a distinct teaching skill; and (2) the activities require the teacher to directly demonstrate some skill which can be assessed by evaluating either the performance or the product, depending on the focus of the activity. One example of the assessment approach was piloted: The Secondary English Assessment. This assessment consisted of three distinct exercises plus a portfolio documenting a unit taught. (The portfolio was discussed earlier as a separate assessment approach.) The three exercises were: Responding to student writing, a "fishbowl" discussion of a literary work, and an impromptu speech on a topic pertaining to language and literature in a multicultural society. The first exercise focuses on subject-specific pedagogical skills, specifically evaluating student learning. The latter two exercises focus on content skills of literary interpretation and speech. The Secondary English Assessment was analyzed in the *Second Year Report.*

Other Assessment Approaches

The last two assessment approaches to be described are distinct both from each other and from the assessment approaches described above. This set of approaches includes videotaped teaching episodes and multiple-choice examinations.

Videotaped teaching episodes. Videotaped teaching episodes require a teacher to respond to questions pertaining to videotaped scenarios of classroom events and activities. Material supplementing the videos, such as copies of stories read by students and a list of questions to be asked after viewing a lesson segment, may be provided. One such instrument, the Language Arts Pedagogical Knowledge Assessment (LAPKA), was pilot tested. The instrument focussed solely on subject-specific pedagogical skills. LAPKA was analyzed in the *Second Year Report.*

Multiple-choice examinations. Multiple-choice examinations require a teacher to answer highly focussed questions about teaching by selecting one or more correct responses from a fixed number of response options. Scoring is typically on a right-wrong basis for each item, though other scoring systems that grant partial credit or deduct for guessing are also used. As multiple-choice examinations are the dominant form of teacher assessment at present, only a single instrument, the Elementary Education Examination, was piloted. The Elementary Education Examination differed from traditional multiple-choice assessments of teaching in that it attempted to embed both theoretical and applied questions in classroom contexts and included some materials-based items. Materials-based items required teachers to read and evaluate documents such as Individual Education Plans, student worksheets, and report cards. The Elementary Education Examination was evaluated in the *First Year Report.*

Data Sources and Analytic Categories

Several sources of data were used for this analysis: Evaluation feedback forms completed by teachers who participated in the pilot tests; evaluation feedback forms completed by the observers and scorers; observations of the administration of each assessment and of the training of observers and scorers as recorded in field notes by FWL staff; scores that reflected the performances of participating teachers on the assessment instruments; review of instruments or portions of instruments by an expert on teaching diverse students; the most recent Curriculum Guide(s) and/or Framework(s); and the performance standards for student teachers in the *Standards for Program Quality and Effectiveness, Factors to Consider and Preconditions in the Evaluation of Professional Teacher Preparation Programs.*

The general analytic categories that were used to appraise the assessment approaches were the same for all instruments. They included an analysis of content, cost, and technical quality. For each individual instrument, these analytic categories are discussed more fully in the two previous reports of the CNTP Assessment Component. In this report, the assessment approaches are compared on the basis of several broad analytic criteria. The first of these is content.

CONTENT ANALYSIS

The ability of the assessment approaches to measure teaching skills were examined along several dimensions: Congruence with State policy documents, measurement of teaching skills, job-relatedness, appropriateness for beginning teachers, and appropriateness across different teaching contexts. Previous reports analyzed individual instruments. This report draws upon those analyses and experiences across all instruments to draw conclusions about the ability of each assessment approach to assess specific domains of a teacher's knowledge and skills.

Congruence of Content with State Policy Documents

Earlier reports examined the congruence of each instrument with two types of State documents: Curriculum Frameworks or Curriculum Guides for the subject area(s) and grade level(s) serving as the focus of the assessment, and the performance standards for student teachers that are contained in each set of program quality standards for credential programs.

The State Curriculum Frameworks for the various subject-matter areas represent a constructivist approach to the curriculum, a departure from previous curricular emphases. At least one Curriculum Framework and/or Model Curriculum Guide has been issued for each subject area as a statement of statewide curricular goals. Therefore, the congruence of each assessment approach with these curriculum documents has been a focus of analysis. While specific assessment instruments exhibited either inconsistencies or omissions when compared with a relevant framework, additions or modifications were easily envisioned which would address the deficiencies without altering the assessment approach. Therefore, congruence with State Curriculum Frameworks seems to be a characteristic of an individual assessment instrument rather than of an assessment approach, as no approach appears to be incompatible with any Curriculum Framework or Model Curriculum Guide.

Each instrument was also analyzed for its ability to assess the California standards of competence and performance for student teachers. These teaching standards were developed as part of a more general reform in the Commission on Teacher Credentialing's approach to evaluating teacher credential programs, which involved issuing program standards which each credential program is expected to meet. These program quality standards include one section which sets forth performance standards for student teachers. The pilot test analysis suggests that not only do individual assessment instruments vary in the ability to measure specific standards, assessment approaches vary as well. Rather than discuss each of the CTC student teacher performance standards separately, the next section discusses the ability of each assessment approach to assess broad domains of teaching.

Measurement of Teaching Skills

Not surprisingly, assessment approaches vary in the degree to which they are able to measure different domains of teaching skills. A draft framework of knowledge, skills, and abilities for beginning teachers produced in another part of the CNTP Assessment Component identifies important teaching domains. Using these categories, this report will focus on each approach's ability to measure five major domains of teaching: Planning and designing instruction, classroom organization and management, instruction, diagnosis and evaluation, and participation in a learning community. Subject-matter knowledge is included as an additional domain, because it was addressed by many of the assessment instruments. (However, assessment of subject-matter knowledge has been the primary focus of several recent reforms that are beyond the scope of the CNTP.) In those cases where the domains were not addressed by the pilot tested instruments, FWL staff drew upon the pilot testing experience with all instruments to try to imagine revisions which might address the unaddressed domain(s). In order to differentiate the strengths and weaknesses of the assessment approaches, ratings are given along three dimensions for each instrument in relation to each domain:

- **Breadth:** The number and range of aspects (e.g., concepts, contexts, situations, skills) that the assessment approach can tap. The key issue is how broadly the approach can sample the domain.

- **Depth:** The extent to which the focal skills, tasks, questions, and/or responses provide extensive evidence of the teacher's knowledge, skill, ability, understanding, and/or reflective reasoning within about the domain.

- **Authenticity:** The degree to which the focal skills, tasks, questions, or responses represent the teacher's thoughts and actions as they occur in actual teaching situations.

The evaluations of each approach are summarized in Table 1. The discussion below iden+ifies the basis for the evaluations, and whether the evaluations are based on pilot test

TABLE 1

ANALYSIS OF ALTERNATIVE ASSESSMENT APPROACHES AND THEIR ABILITY TO ASSESS SPECIFIC TEACHING DOMAINS

TEACHING DOMAINS	CLASSROOM-BASED ASSESSMENT APPROACHES					
	Classroom Observations			Portfolios		
	Breadth	Depth	Authenticity	Breadth	Depth	Authenticity
Planning and Designing Instruction	⊙	⊙	◒	◒	●	●
Classroom Organization and Management	●[1]	●[1]	●	⊙	⊙	⊙
Instruction	●[1]	●[1]	●	⊙	⊙	⊙
Diagnosing and Evaluating Student Learning	●[1,2]	●[1,2]	●[1,2]	◒	◒	●
Participating in a Learning Community	⊙[3]	⊙[3]	⊙[3]	○	○	○
Subject Matter Knowledge	⊙	⊙[3]	◒[3]	⊙	◒	◒

Extent of Coverage

●	Extensive
◒	Moderate
⊙	Some
○	None

[1] If more than one observation, and lesson characteristics vary
[2] Ratings are mainly based on the monitoring and adjusting of instruction
[3] Depends on the lesson and lesson objectives
[4] Varies with skill and/or area of knowledge
[5] For subject areas with a visual performance component, the ratings would be moderate for breadth, and extensive for depth and authenticity

TABLE 1 (Continued)

TEACHING DOMAINS	PERFORMANCE SIMULATIONS APPROACHES								
	Semi-Structured Interviews			Structured Simulation Tasks			Performance-Based Assignment Center Exercises		
	Breadth	Depth	Authenticity	Breadth	Depth	Authenticity	Breadth	Depth	Authenticity
Planning and Designing Instruction	⊙	●	◒	◒	◒	◒	◒	◒	◒
Classroom Organization and Management	◒	◒	⊙	◒	⊙	⊙	◒	◒	⊙
Instruction	⊙	◒	⊙	◒	⊙	⊙	◒	◒	⊙
Diagnosing and Evaluating Student Learning	◒	●	◒	◒	⊙	◒	◒	●	◒
Participating in a Learning Community	⊙	◒	○	⊙	⊙	⊙	⊙	◒	○
Subject Matter Knowledge	⊙	●	◒	⊙	⊙	◒	⊙	●	●[4]

Extent of Coverage

●	Extensive
◒	Moderate
⊙	Some
○	None

[1] If more than one observation, and lesson characteristics vary

[2] Ratings are mainly based on the monitoring and adjusting of instruction

[3] Depends on the lesson and lesson objectives

[4] Varies with skill and/or area of knowledge

[5] For subject areas with a visual performance component, the ratings would be moderate for breadth, and extensive for depth and authenticity

TABLE 1 (Continued)

TEACHING DOMAINS	OTHER ASSESSMENTS					
	Videotaped Teaching Episodes			Multiple-Choice Examinations		
	Breadth	Depth	Authenticity	Breadth	Depth	Authenticity
Planning and Designing Instruction	⊙	◓	⊙	◓	⊙	⊙
Classroom Organization and Management	◓	◓	⊙	●	⊙	○
Instruction	◓	◓	⊙	◓	⊙	○
Diagnosing and Evaluating Student Learning	⊙[5]	○[5]	○[5]	◓	⊙	⊙
Participating in a Learning Community	○	○	○	⊙	⊙	○
Subject Matter Knowledge	⊙[5]	⊙[5]	⊙[5]	●	⊙	⊙

Extent of Coverage

●	Extensive
◓	Moderate
⊙	Some
○	None

[1] If more than one observation, and lesson characteristics vary

[2] Ratings are mainly based on the monitoring and adjusting of instruction

[3] Depends on the lesson and lesson objectives

[4] Varies with skill and/or area of knowledge

[5] For subject areas with a visual performance component, the ratings would be moderate for breadth, and extensive for depth and authenticity

experience or inferences about hypothetical revisions. The discussion of each assessment approach occurs in the same order in which they were previously described, and begins with high-inference classroom observations.

High-inference Classroom Observations

Of the two high-inference classroom observation instruments piloted, one focussed on general pedagogy and the other examined both general and subject-specific pedagogy in the area of laboratory science. In general, the following conclusions can be made about the ability of the high-inference classroom observation approach to measure teaching skills:

- The strength of high-inference classroom observations lies primarily in the authenticity dimension, as this approach accurately reflects a teacher's actions as they occur in actual teaching situations.

- The approach allows for some depth in the assessment of classroom organization and management, instruction, and the monitoring and adjusting of instruction.

- High-inference classroom observations are extremely weak in sampling ability, however, as the assessment approach only measures what is seen at a particular point in time in a particular setting.

With respect to designing and planning a lesson, the two instruments piloted required teachers to provide the observer with an outline of the lesson design ahead of time, so that observers could understand the lesson goals and the sequence of activities which they were likely to observe. High-inference classroom observations provide evidence regarding some of a teacher's knowledge and skills in planning and designing instruction. An observer can see whether any materials needed are readily accessible, whether activities planned seem to be appropriate for the students, and whether the teacher has anticipated the cognitive and behavioral responses of the students to the lesson. Lessons aimed at higher-order thinking can also provide instances of planning to promote the development of these skills. The observer may be able to note if the lesson draws upon student interests or upon relevant aspects of the students' lives or from the community in which the school is located, which is particularly important in teaching diverse students. An observer cannot evaluate how well the lesson observed builds upon previous student learning and contributes toward meeting more global objectives, if any, for the unit or school year. A single lesson, or even a set of unrelated single lessons, is unlikely to provide opportunities to display a broad range of planning skills. Moreover, the alternative activities considered and the reasons underlying the choice of the activities observed are not apparent through observations alone. This lack of information about a teacher's rationale, and the difficulty of making valid inferences about a teacher's planning skills from the activities observed, limits a high-inference classroom observation's ability to assess the depth of a teacher's skills in this domain. While the classroom provides an ideal site to evaluate the teacher's ability to apply instructional plans, the inability to observe many of a teacher's thoughts and actions while planning means that the authenticity with which planning is evaluated is moderate.

High-inference classroom observations provide better coverage of the skills in classroom organization and management, yielding rich information about the management of student behavior, classroom organization, and the degree to which students are actively involved in instructional activities. The ability to evaluate a teacher's skill at fostering independent learning by students is dependent on the lesson goals of the lesson(s) observed. Observation of multiple lessons portraying a range of patterns of classroom organization (whole group, small group, individual) provides extensive information about a broad range of skills in organization and management. A teacher's behaviors across multiple observations can also form the basis of inferences about the depth of their skills in this domain. Since most of a teacher's skills in classroom organization and management are related to their interaction with students, an observation is the single best assessment approach to gather extensive authentic evidence about a teacher's organization and management skills.

Since interaction with students is also at the heart of the instructional domain, high-inference classroom observations are uniquely suited to gathering evidence regarding the breadth and depth of skills in this domain such as communicating with students, motivating students and engaging them in appropriate instructional activities. The breadth of skills observed in individual lessons is limited by the type of students, topic, and learning objectives of the lesson observed. However, observations of multiple lessons which vary along these dimensions can provide extensive opportunities to gather evidence of the breadth and depth of a teacher's instructional skills. The ability of high-inference observations (i.e., observations which require professional judgment for evaluation and are not just checklists of behaviors to be observed) to collect extensive information about authentic teacher-student interactions related to instruction makes this the best-suited assessment approach to assess instructional skills.

High-inference classroom observations are also uniquely suited to measuring one set of skills in the domain of diagnosing and evaluating student learning: Monitoring and adjusting instruction during a lesson. Observations of multiple lessons requiring varying monitoring skills and strategies (e.g., reviewing familiar material, practicing skills already learned, initially practicing skills) can provide evidence regarding an extensive array of skills relating to monitoring and adjusting instruction during a lesson. Extensive evidence can also be collected across observations which gives an idea of the depth of skills, e.g., the subtlety of student responses indicating either understanding or confusion to which a teacher responds. As with the other domains where teacher-student interaction is a prominent feature, high-inference classroom observations collect highly authentic evidence of a teacher's skills in monitoring and adjusting instruction during a lesson. However, high-inference classroom observations are more limited in their ability to evaluate other skills in this domain, such as a teacher's summative evaluation strategies, methods of communicating progress to students and their parents, and diagnostic strategies beyond monitoring and adjusting instruction during a lesson. High-inference classroom observations can provide some indication of the breadth and depth of these other evaluative skills if the teacher provides oral or written summative feedback to the students during class, e.g., in individual conferences or in small groups. However, these type of activities typically do not occur frequently during lessons. Since high-inference classroom observations cannot completely reflect teacher thoughts and behaviors that occur during more summative evaluations, they provide only a moderate degree of authenticity with respect to reflecting this category of skills and abilities in this domain.

The next domain is participation in a learning community, which includes keeping abreast of current developments in the subject-matter area and grade level, and making appropriate use of school- and community-based resources. With the exception of evaluating the use of an aide, if present, no piloted observation instrument attempted to examine this dimension. Activities in this domain typically occur outside of the classroom. Opportunities to observe evidence with respect to this domain are highly variable, as some lesson activities draw upon school and community resources more than others. Therefore, at best, high-inference classroom observations can only provide some indication of the breadth of skills in this domain. Because the observer typically is not familiar with the resources available to the teacher, only some indication of the depth of knowledge, skills, and abilities in this domain is possible from the use or lack of use of school and community resources observed in the classroom. Since most activities within this domain occur outside the context of classroom instruction, only some authentic activities in this domain can be captured by high-inference classroom observations.

Subject-matter knowledge can be observed in a lesson. Sometimes the subject matter is explicit, such as when the teacher defines concepts, explains processes, or responds to questions. Other times, however, subject-matter knowledge is more implicit, evidenced in the questions and answers provided by the teacher or in the lesson design. Egregious errors in content can be identified. However, even a set of several lessons spaced over time will not reliably provide much indication of breadth of subject-matter knowledge, as the number of possible topics is too large to be covered by a small number of observations. The depth with which observations tap subject-matter knowledge depends on the lesson and the lesson objective. The more complex the higher-order thinking skills called for in the lesson, the more likely that a teacher is called upon to display some depth of subject-matter knowledge. However, initial explanations of concepts and processes which rely on metaphors are particularly difficult to evaluate, because the metaphor almost always coveys at least some inaccuracies about content. Even opportunities for the unambiguous display of more complex subject-matter knowledge occurs in the context of simultaneous management of student behavior, progress through the lesson, and quick responses to student questions. It is possible that a teacher has far more depth of subject-matter knowledge than is evidenced in the classroom, so high-inference classroom observations can only provide some coverage of the depth of a teacher's subject-matter knowledge. Observers have opportunities to see how subject-matter knowledge is applied in actual teaching situations. However, while the subject-matter knowledge displayed to the observer has high authenticity because it occurs in the context of teaching, the observation of errors cannot always be interpreted to mean lack of subject-matter knowledge. At best, observations can provide a moderate amount of authentic evidence about subject-matter knowledge.

Portfolios

One portfolio instrument was piloted in conjunction with the Secondary English Assessment. This portfolio documented the teaching of a unit of instruction, although portfolios could instead contain examples of what the teacher considers their best work in specific categories. Documentation of a unit works well because a portfolio provides

multiple, related sources of evidence regarding the teacher's skills. In general, the strengths and weaknesses of portfolios in measuring teaching skills are as follows:

- Portfolios are strongest with respect to the authenticity with which they measure skills in the domains of planning and designing instruction, diagnosis and evaluation of student learning, and, to a lesser extent, subject-matter knowledge.

- Portfolios are weakest in the ability to assess skills in the domains of classroom instruction, organization, and management.

- The ability of portfolios to measure depth of knowledge varies. They are strongest in measuring the depth of skills in planning and designing instruction, and weakest in measuring skills related to teacher-student interaction.

If constructed to document the teaching of a unit, a portfolio can require extensive documentation of the planning and design of the unit, including a unit plan, materials and assignments given to students, samples of student work, a log of significant events and/or insights and resulting changes in plans, and a self-reflective essay at the end of the unit summarizing what the teacher has learned from the experience of teaching the unit. Teachers can also be required to state their unit objectives and give a brief description of their classroom context that affects how they construct and teach the unit. In the piloted instrument, open-ended student evaluations of the unit taught, responding to questions such as "What did you enjoy most about the unit?" and "What changes do you suggest for the next time the unit is taught?" often suggested problems in the teaching of the unit to which the teacher might have responded. If all of these elements are included in a portfolio, there is evidence on virtually every facet of the planning and design of the unit. If the unit is chosen by the teacher, then the portfolio should represent their best efforts within the curriculum they are actually teaching. However, since the information only addresses the teaching of a single topic, then the portfolio can only partially measure the breadth of a teacher's skills in planning and designing instruction. Teacher commentary and self-reflection on the unit taught, accompanied by student work or evaluations which give independent evidence of factors which might have affected planning, provide rich information to extensively evaluate the depth of a teacher's thinking about planning and designing instruction. Since the portfolio documents a unit actually taught, it is high in the authenticity of the skills assessed in planning and designing instruction.

Portfolios provide a small degree of coverage of all dimensions of a teacher's classroom organization and management skills. The best evidence in this domain is contained in a list of unit activities, which can indicate the extent to which a teacher actively involves students, fosters independent learning, and utilizes different student grouping patterns. Evidence about behavior management is unlikely to appear in the portfolio materials unless the teacher is experiencing problems which interfere with teaching the unit, in which case a reflective essay or weekly log might document the teacher's struggle. It is unlikely that additional portfolio components could address skills in this domain. The teacher might be asked to more fully describe the rationale for classroom groupings and to discuss their behavior management strategies as part of the contextual information, and to include management strategies in their lesson outlines. However, these additional components would only marginally improve the available evidence.

In the domain of instruction, a portfolio again provides a small amount of evidence addressing each dimension. Some information about a teacher's communication skills can be obtained through the quality of the written handouts given to students. The introductory activities, both for each lesson and for the unit as a whole, give some insight into a teacher's motivational strategies. The extent to which different activities are planned throughout the unit or within a single lesson together with the description of students provided by the teacher can give a sense of whether the teacher is accommodating individual differences in achievement, interests, backgrounds, and learning styles. However, as with other assessment approaches which do not observe a teacher's interaction with students in the classroom, the portfolio misses the interactional aspects which are central to instructional skills, severely limiting the ability to assess the breadth and depth of skills in this domain. Although the components described above capture some authentic aspects of instruction, it is difficult to think of additional components which could be added to a portfolio to more fully capture skills in this domain.

A portfolio is more successful at documenting some aspects of a teacher's diagnosis and evaluation of student learning, if a component centering around the evaluation of student learning can be included. The component could begin with a brief outline of evaluation strategies used during the unit, with more elaborated descriptions of methods that are not apparent from a document, such as evaluating oral presentations. The teacher can also be asked to include samples of all student work products, observation notes, and other forms of documentation of student learning used to evaluate students. The pilot tested portfolio asked teachers to provide the work of two students at varying levels of achievement across the entire unit, and to provide the work of the entire class for one assignment. All work samples were to include teacher evaluations and comments, providing some evidence on how the teacher communicated progress to students. Evidence of how the teacher adjusted instruction was possible through the weekly logs and the self-reflective essay at the end of the unit. Student evaluations sometimes suggested difficulties in learning, such as unfamiliar vocabulary and archaic speech which made it hard to understand a specific short story. However, routine monitoring and adjusting in the course of a lesson does not generally appear in documentation, and thus this area of competence in the domain of diagnosis and evaluation of student learning is difficult to assess through a portfolio. Therefore, the portfolio can only be said to moderately capture the breadth of skills in this domain. It is better in terms of the depth of teacher thinking portrayed, as the weekly log and the self-reflective essay at the end of the unit can provide extensive evidence as to how the teacher was (or was not) using the evaluation information to make changes in the unit or lessons. Although there is little or no information on routine monitoring and adjusting of instruction, the portfolio documentation could provide extensive examples of many aspects of the diagnosis and evaluation of student learning, making it high in authenticity.

The documentation required in a portfolio does not routinely provide any evidence on a teacher's skills in participating in or creating a learning community. Furthermore, it is difficult to think of any additional documentation related to the teaching of a unit that could reliably be expected to produce such evidence. Teachers with students with special needs who interact with specialists could provide additional documentation of their coordination with those specialists, but this would not apply to every teacher. Some use of non-school resources in the unit might be evidenced, but it is not likely that this would be true of every unit. Therefore, portfolios have virtually no ability to evaluate a teacher's skills along any dimension of this domain.

Subject-matter knowledge is not a separate entry in a portfolio, but is implicit in many forms of documentation which could be included in a portfolio such as the choice and sequencing of activities in a unit plan, definition and/or use of concepts in handouts, responses to student work, and any essay or log containing teacher reflections. These types of documentation could give some evidence of a teacher's knowledge of key concepts and processes, alternative methods of representing or communicating concepts or ideas, ways of judging or valuing products or ideas, and perhaps the use of concrete and applied examples of concepts or integrating ideas, information, or applications from a different subject area. Because a portfolio is developed over a limited period of time, it is unlikely that it could contain many topics; hence its ability to sample a teacher's subject-matter knowledge is not broad. In terms of depth, the structuring and teaching of a unit coupled with some requirement of demonstration of reflection can give some sense of the depth of subject-matter knowledge upon which a teacher draws. Since the focus of a portfolio is on documenting pedagogy and resultant student learning, however, the evidence will be more implicit than explicit unless the teacher is asked to explain their instructional goals and situate them within both previous and subsequent curriculum content (which would be difficult for most beginning teachers, who are only beginning to encounter the curriculum as practiced in their district.) Thus, portfolios can only partially evaluate the depth of a teacher's subject-matter knowledge and skills. In terms of authenticity, the portfolio should provide numerous examples of the application of subject-matter knowledge in the forms of documentation described above, especially if samples of student work with teacher responses are included. However, teachers also apply their subject-matter knowledge in numerous other ways which cannot be captured in a portfolio, such as quickly responding to student questions, so portfolios can only partially capture a teacher's ability to apply subject-matter knowledge.

Semi-structured Interviews

Three semi-structured interviews representing two subject-matter areas and two grade levels (elementary and secondary) were evaluated. Each interview was centered around one to two topics within the subject-matter area. Strengths and weaknesses of semi-structured interviews in measuring teaching skills are as follows:

- The strength of semi-structured interviews is the depth of evidence collected regarding a few skills or topics, as the interview format requires teachers to provide rationales for their decisions. This information is not collected by many assessment approaches, and it which sometimes significantly impacts the evaluation of the teacher's performance.

- Semi-structured interviews are weakest in representing the breadth of a teacher's knowledge and skills, as they cover only a small number of topics.

- The degree to which semi-structured interviews authentically represent a teacher's knowledge and skills is highly variable across domains, and the relationship of a teacher's verbal responses to actual teaching is unknown.

Semi-structured interviews have a limited ability to sample a broad range of skills in planning and designing instruction. Two to three tasks per interview centering around planning and designing instruction (e.g., planning a lesson or structuring a unit by grouping and sequencing provided topics) were piloted. These tasks addressed several competencies in this domain, including sequencing instruction, building on student resources, and the focussing on higher-order thinking skills in planned instruction. Planning skills, however, are likely to vary with the familiarity of the topic, and each interview only covers a small range of topics. In addition, the skill of evaluating and adapting plans either as or after they are implemented is difficult to address through simulated tasks. In contrast to the breadth of coverage, semi-structured interviews can extensively evaluate the depth of a teacher's planning skills. The interviews in social studies were much richer than the interviews in mathematics in terms of teacher descriptions of how they drew upon student backgrounds and interests, probably reflecting differences between the social sciences/humanities and the sciences in the degree to which the impact of student backgrounds and interests upon instruction is apparent. The interview format requires teachers to describe the rationale underlying their hypothetical instructional decisions, which provides an opportunity for teachers to display the depth of their planning skills. The authenticity of the skills evaluated by semi-structured interviews is moderate. While tasks can require teachers to actually outline instructional plans for hypothetical students and describe the factors taken into consideration in the design, they cannot measure teachers' skills in actually making those decisions in their own teaching context under the time pressures faced by classroom teachers.

In terms of classroom organization and management, the only tasks that addressed this domain in the pilot studies were those related to lesson planning. The extent to which students were actively involved and the fostering of independent learning was evident in teachers' lesson plans and descriptions of activities chosen. Tasks that are more directly focussed on this domain could be envisioned, however. Such tasks might require a teacher to describe a behavior management system and/or describe the strengths and weaknesses of various ways of organizing the classroom. Again, teachers can be asked to describe their rationales for their decisions. The competencies in classroom organization and management are less likely to vary with the subject-matter topic, and more likely to vary with a relatively small number of types of instructional goals. Therefore, a moderate breadth of skills in classroom organization and management can be assessed by the interview format. Much of a teacher's thinking in this domain occurs in the form of simultaneously processing information about student behavior, motivation, and progress toward instructional goals. This type of thinking is extremely difficult to cue outside of the classroom, and hence difficult to elicit through an interview. Therefore, semi-structured interviews can only moderately evaluate the depth of a teacher's thinking about classroom organization and management. A teacher's ability to describe a rationale for classroom organizational and management decisions may vary considerably from the ability to actually apply these skills where plans may easily go awry and the implementation of these decisions require constant monitoring of multiple sources of information. Semi-structured interviews provide evidence of how teachers think about such decisions, but not about the ability to actually implement them, so such interviews provide limited information about applied skills regarding classroom organization and management.

Semi-structured interviews also have a limited ability to assess a teacher's instructional skills. Tasks addressing this domain included instructional vignettes in which the teacher was to respond to hypothetical student instructional problems. None of the tasks explored linguistic diversity in the classroom, though this, too, could be the focus of a specific question or task. The interviews could also be modified to ask teachers to describe how they approach student misunderstandings, motivation, and individual and group differences. Many interpersonal skills such as reading students' nonverbal cues, soliciting and using information about student backgrounds and interests during instruction, and interactional skills with children and youth can only be indirectly reflected, at best, by the performance of tasks. If teachers perform a task related to instruction, e.g., responding to instructional vignettes, then they can explain the rationale underlying their decisions, and the depth of their thinking can be assessed. However, during instruction, teachers also react to numerous cues, such as nonverbal student behavior, which cannot be accurately conveyed through simulations. They also have an opportunity to elicit information and test out hypotheses through posing questions to and engaging in interaction with students. Thus important clues available in the classroom are lacking in simulations, so semi-structured interviews are only able to moderately evaluate the breadth and depth of a teacher's instructional knowledge and skills. These factors also severely limit the ability of semi-structured interviews to assess the ability of teachers to apply their knowledge of instruction in the classroom.

Several tasks were piloted which specifically addressed the domain of diagnosis and evaluation of student learning, including evaluating student work, designing a quiz, explaining how monitoring of student understanding was to occur within a lesson plan, and planning remediation for a student exhibiting repeated misunderstanding of a central concept on a simulated test. An additional task or task component asking the teacher to describe strengths and weaknesses of various approaches to summative evaluation could also be designed. However, skills of quickly identifying student misunderstandings during the course of a lesson and attempting to address them, as well as communicating student progress to students, parents, and resource specialists are not easily measured through simulated activities. Therefore, semi-structured interviews can only moderately address a teacher's skills in diagnosis and evaluation of student learning. Teachers can extensively describe their strategies of diagnosing and evaluating student learning to illustrate depth. However, because teachers know much more about their own students (e.g., temperament, previous learning) than can ever be conveyed in a simulation, the display of responses to hypothetical student work can only moderately reflect teachers' application of those skills in actual classrooms.

A few piloted tasks addressed the domain of participating in or creating a learning community by including non-school resources (e.g., the local library, an amateur historian well known in the community, parent volunteers) among potential pedagogical approaches to be critiqued, and teachers could use similar resources in designing lesson plans. A task could also be created which probes the depth of a teacher's knowledge of the use of hypothetical school and non-school resources. Many of the interactional skills in this category, such as effectively working with other adults, evaluating the strengths and weaknesses of resource specialists with whom one works, and developing effective networks for learning about resources in the school, community, and profession, are difficult to simulate. Therefore, semi-structured interviews have a very limited ability to reflect the breadth of a teacher's ability to participate in or create a learning community. Though it is likely that many beginning teachers have not had enough experience to develop complex strategies, beginning teachers could describe or critique strategies related to a learning community (e.g., learning about community resources, working with specialists). However, a beginning teacher's knowledge of learning communities is likely to be strongly grounded to their particular teaching context, and interactional skills will strongly affect a teacher's skills with respect to participating in a learning community. A task would have to be carefully designed to enable a teacher to display a moderate amount of the depth of their knowledge and skills in this domain. As very little evidence apart from a teacher's verbal description of their strategies for participating in or creating a learning community is available, semi-structured interviews provide only a limited reflection of a teacher's thoughts and actions when applying their knowledge and skills in this domain.

While subject-matter knowledge was at least implicit in every task that was pilot tested, some tasks, such as unit planning or topic sequencing, lesson planning, evaluating alternative mathematical approaches, and evaluating student learning, provided extensive information about subject-matter knowledge of a particular topic. Because a reasonable number of tasks can cover only a small number of topics, however, the breadth of subject-matter knowledge portrayed is very limited. The developers of the piloted interviews chose to focus the interviews on topics which were either widely taught (e.g., the Civil War) or which represented more fundamental knowledge in a more hierarchically organized subject area (e.g., linear equations, linear functions, fractions, ratios and proportions). The set of questions and the request to fully explain the rationale underlying judgments in a semi-structured interview provide teachers with opportunities to exhibit a great depth of subject-matter knowledge. The resemblance of performance in the interview to actual application of subject-matter skills is, however, moderate at best. While some of the subject-matter analysis, e.g., that in lesson planning, resembles what a teacher does, the pressures of time and the demands of classroom management may reduce the depth with which a teacher applies their subject-matter knowledge in actual practice.

Structured Simulation Tasks

Three sets of structured simulation tasks were piloted. In general:

- The strengths of structured simulation tasks lie in their ability to measure the application of skills in the areas of subject-matter knowledge, planning and designing instruction, and diagnosing and evaluating student learning. However, the relationship of these hypothetical responses to actual teaching decisions is unknown.

- The weakness of structured simulation tasks is in the lack of depth with which teaching skills are measured.

- The ability of structured simulation tasks to sample teaching skills broadly is limited by the number of tasks, many of which portray a single topic, instructional strategy, and teaching context.

In the domain of planning and designing of instruction, tasks included critiquing lesson plans, outlining lessons or lesson sequences, assembling a unit plan from components provided to address specific goals for a specific classroom of students, and critiquing transcripts of simulated lesson segments. These tasks addressed most aspects of planning and design of instruction, although for a limited number of topics, so a moderate degree of breadth of skills is reflected. Structured simulation tasks also moderately reflect the depth of a teacher's planning skills. Although the underlying rationale was not solicited, some indication of the complexity of a teacher's skills in planning and instructional design was evident in the teacher responses. This assessment approach simultaneously assesses a teacher's planning skills, subject-matter knowledge, and knowledge of principles of effective instruction of specific types of students. However, the students are only described along a few dimensions, and the rich contextual information and individual variation present in classrooms is lacking. While structured simulation tasks can capture a teacher's knowledge of general principles on instructing a variety of types of students, they cannot capture a teacher's ability to capitalize upon personal knowledge of students and a specific context. Furthermore, the ability to display planning skills can be constrained by inadequate knowledge of the particular subject matter, type of student specified, or, sometimes, specific instructional techniques. Therefore, structured simulation tasks can only moderately reflect a teacher's application of planning skills.

In the domain of classroom organization and management, simulated tasks have difficulty in capturing a teacher's ability to manage student behavior, since most of these skills are interactional and highly idiosyncratic in nature (with respect to both the teacher and with respect to the students). Behavior management and classroom organization were, however, addressed through analysis of simulated classroom transcripts. Identifying strengths and weaknesses of specific features of classroom organization evident in the transcript and the ways in which a particular organizational scheme were implemented seemed to work well in assessing a teacher's knowledge of classroom organization. In contrast, it seemed difficult to illustrate behavior management through transcripts, particularly strengths in behavior management, beyond a superficial level. Tasks requiring lesson designs provided opportunities to see how actively the teacher's design involved students and fostered independent learning, aspects of classroom organization and management. Because much of classroom organization and management is interactional in nature and therefore difficult to simulate, structured simulated tasks can only capture a limited breadth of skills in this domain. The response format provides some indication of the complexity of a teacher's thinking through the complexity of the cues to which they respond, but the lack of opportunity for a teacher to explain the rationale underlying decisions or evaluations limits the ability to reflect the depth of a teacher's thinking. As many of the thoughts and behaviors in this domain depend on interaction and are beyond the scope of simulation, structured simulated tasks only provide some authentic representation of a teacher's skills in classroom organization and management.

Piloted tasks which addressed the domain of instruction included critiquing simulated classroom transcripts in terms of subject-specific pedagogical skills and communication with students, choosing activities which are likely to stimulate student interests, and responding to problems in vignettes representing instructional decisions. However, interactional aspects of these skills were not addressed, and are difficult to simulate. Thus structured simulation tasks can only moderately represent knowledge and skills in the instructional domain. Decisions about instruction require much tailoring to the instructional context in which a teacher works, which is extremely difficult to simulate, and the response format does not elicit the rationale underlying judgments. Therefore, structured simulation tasks are very limited in reflecting the depth of a teacher's thinking. While responding to instructional vignettes simulates some types of instructional thinking, responding to transcripts of classroom interaction is very remote from the type of instantaneous response which occurs in the classroom as the teacher responds to students. Therefore, the authenticity with which the tasks reflect instructional skills is also very limited.

Tasks focussing on the diagnosis and evaluation of student learning included critiquing a transcript of a simulated lesson where the teacher responded to students giving correct and incorrect responses, reviewing a transcript where students are exhibiting misunderstanding and designing a subsequent activity aimed at remediation, and evaluating simulated pieces of student writing. It is well within the assessment methodology to envision a task which asks a teacher to list strengths and weaknesses of various evaluation strategies for assessing whether students are meeting specific instructional goals. Another possible task would be the construction of a test or quiz, given some indication of content that was taught. However, it would be difficult to construct a task beyond transcript review that is centered around routine monitoring and adjusting of instruction, an important competency in this domain. Therefore, like semi-structured interviews, structured simulation tasks are judged to only moderately reflect the breadth of a teacher's skills in diagnosis and evaluation of student learning. The tasks either piloted or envisioned can illustrate some complexity in teacher thinking about summative evaluation and the design of remedial instruction. However, within this methodology, teachers do not offer rationales for their evaluation decisions. In semi-structured interviews, the presence of such a rationale lent certainty to what would otherwise be inferences from little data about a teacher's decisions in this domain. Therefore, structured simulation tasks can only reflect the depth of a teacher's knowledge in diagnosing and evaluating student learning in a very limited way. The type of response that teachers make to the tasks, especially in responding to simulated student work, closely resembles the thinking processes employed in actual classroom situations. As with other simulations, it is difficult to simulate the routine monitoring and adjusting of instruction in response to indications of student understanding. Therefore, structured simulation tasks can only moderately reflect authentic thoughts and decisions in this domain.

In the domain of participating in or creating a learning community, the only task which addressed this set of skills focussed on the creation of a parent/student letter describing a specific course to acquaint parents and students with the usefulness of the course and to identify course requirements and teacher expectations. Other than creating a task centered around interaction with a school specialist or other teacher, or incorporating a few elements of utilization of school or non-school resources into a lesson design, it is difficult to envision additional simulation tasks which directly address this domain. Therefore, only some skills in this domain can be portrayed by simulation tasks. The difficulty of portraying these skills in written responses to a simulation task, and the methodology of limiting the scope of the response to facilitate scoring means that only a very limited depth of knowledge in this domain can be represented. Both contextual information and interaction with others are central to skills in this domain. These can only be authentically reflected in simulation tasks in very limited ways.

Structured simulation tasks have a very limited ability to sample the breadth of a teacher's subject-matter knowledge. While no task piloted centered around subject-matter knowledge, such knowledge is embedded within the stimulus materials and teacher responses. Tasks often contain cues that are apparent only to those with appropriate subject-matter knowledge, and responses often depend upon subject-matter knowledge for their successful construction. One task in science, for example, required teachers to use their knowledge of chemistry to identify safety hazards in a picture of a storage cabinet. To respond appropriately to tasks, teachers needed to know key concepts and sometimes their interrelationships, know alternative methods of representing or communicating concepts, identify the use of concrete and applied examples or activities as a strength of a lesson or plan, and use appropriate techniques to judge the appropriateness of a student response. Each task could potentially represent a different topic, but the number of tasks possible is very limited, compared to the number of potential topics. Thus, only some breadth of subject-matter knowledge can be represented. In terms of depth, the knowledge that the teacher is asked to display varies in terms of complexity, but the response itself is usually quite limited. A short response to facilitate scoring is encouraged, often taking the form of a list or an outline. Responses usually display subject-matter knowledge, subject-specific pedagogy, and knowledge of students in an integrated performance. While the presence of a correct response may demonstrate complex subject-matter knowledge, the lack of a correct response may be due to lack of knowledge of pedagogy, lack of knowledge of students, or to overlooking an important cue in the stimulus materials, and does not necessary mean lack of depth of subject-matter knowledge. Thus, structured simulation tasks have only some ability to measure the depth of subject-matter knowledge. In terms of application, the ways in which teachers are asked to display their subject-matter knowledge resemble the use of subject-matter knowledge in teaching responsibilities. However, structured simulation tasks cannot simulate the application of subject-matter knowledge in interchanges between teachers and students. Therefore, semi-structured interviews only partially reflect the authentic display of subject-matter knowledge in teaching.

Performance-based Assessment Center Exercises

In a sense, the content of performance-based assessment center exercises is the easiest to evaluate. Each exercise focusses on a different teaching skill and is independently administered, so each can use a different assessment approach (other than classroom-based approaches). Other factors, such as costs or other foci, may affect the choice, but Table 1 provides guidance as to the assessment approaches which can be administered in an assessment center that best assess each domain.

For knowledge and skills in planning and designing instruction, structured simulation tasks appear to be the best approach to be used for a performance-based assessment center exercise. For classroom organization and management, semi-structured interviews or videotaped teaching episodes (described later) seem to be most useful. Videotaped teaching episodes are also the most powerful simulation approach to measure instructional skills. Diagnosis and evaluation of student learning seem to be best measured by a semi-structured interview approach. No good performance-based approaches were identified to measure skills in participating in or creating a learning community.

None of the assessment approaches described in this section focussed solely on the assessment of subject-matter knowledge. Two of the three exercises piloted in the performance-based assessment center instrument, however, focussed almost solely on specific competencies within English subject-matter knowledge, and the third exercise combined subject-matter knowledge with content pedagogy. Because of the number of competencies and topics in subject-matter knowledge, it is not possible to portray more than a limited amount of breadth through exercises, as each exercise can only focus on a minimal number of competencies and topics. The depth of a teacher's knowledge which can be displayed, however, can be extensive. The extent to which the skills and knowledge demonstrated resemble those applied in the course of teaching varied considerably for the pilot test exercises, ranging from hardly at all for one of the exercises addressing skills in delivering a speech to extensively for an exercise in evaluating student writing.

Videotaped Teaching Episodes

Videotaped teaching episodes were the most difficult assessment approach to analyze, because there was only one prototype developed, and it did not seem to fully capitalize on the video medium. In particular, the short question/short response format which focussed a teacher's attention on discrete ideas or bits of information did not seem to capitalize on the video stimulus. Other response formats such as those used for other assessment approaches might make more use of the information contained in a videotape. Most other assessment approaches were able to build on previous experience with the approach, either within or outside of education. However, videotaped teaching episodes are breaking new ground, and more problems remain to be solved than was the case for other assessment approaches. Therefore, unlike discussions of other assessment approaches, the discussion of videotaped teaching episodes will not rely heavily on pilot testing experience, but will be more speculative.

Videotapes were used to train scorers for some of the other assessment approaches. Observations of the use of those videotapes suggest both potential uses and potential problems in using videotaped teaching episodes to stimulate teacher responses. Examples of potential uses will be discussed below with respect to specific domains of teaching. The problems are more general, however. Videotaping presents technical problems. While the videotapes used in the assessment prototype were with one exception of extremely high technical quality, the teaching episodes portrayed were limited to individuals or small groups. It is extremely difficult to portray whole class instruction where both the teacher and the students are both visible and audible. Switching back and forth between the teacher and students sometimes misses important cues about teacher or student behavior. Because of these unsolved problems with the videotaped teaching episodes approach, it must be emphasized that conclusions for this assessment approach are therefore extremely tentative.

It is difficult to imagine how planning and designing instruction could be portrayed assessed using a videotape stimulus except by showing a lesson where the students are generally showing confusion or misunderstanding, then asking a teacher to design a subsequent lesson aimed at remediation. However, this is a relatively small aspect of planning and designing instruction. It would be more difficult to provide cues via a videotape that would yield evidence of the type of planning skills evident in unit planning, building on student cultural/linguistic backgrounds, interests, and previous achievement. Therefore, videotaped teaching episodes seem very limited in the ability to assess the breadth of a teacher's knowledge about planning and designing instruction. The depth with which the approach can measure a teacher's skills in this domain depends on the type of response required. The most complex response to a video stimulus would probably be to ask the teacher to produce a plan in response to what was observed, together with an explanation of the rationale for the plan. Like structured simulation tasks, a video stimulus would fail to capture the extent to which a teacher considers their previous experience with students and their previous learning in planning, suggesting that videotaped teaching episodes can only moderately reflect the depth of a teacher's skills in planning and designing instruction. The authenticity of the planning and designing skills in such a plan reflects the thinking processes used in actual teaching only to a limited extent, because in actual teaching, the teacher draws upon more knowledge of particular students, their previous responses to instruction, and on curricular goals when planning instruction.

Videotaped teaching episodes can portray a variety of styles of classroom organization and management, asking teachers to identify the important features of the methods used and to critique their appropriateness and/or effectiveness for both the lesson observed and for other types of lessons. Thus videotaped teaching episodes could moderately reflect the breadth of a teacher's knowledge of classroom organization and management techniques. Many interactional competencies within behavior management are contingent upon knowledge of individual students. This information cannot be communicated through a video stimulus, and can be duplicated only in a limited manner through supplementary descriptions. If a critique or extension of the techniques observed is included as a part of the response, then videotaped teaching episodes can assess the depth of a teacher's thinking about classroom organization and management to a moderate degree. The type of thinking displayed, however, resembles that which occurs in actual teaching situations in only very limited ways, as classroom management requires many decisions which occur during instruction and require a constant balancing which includes maintaining progress toward achieving instructional goals, anticipating student behavior, and monitoring student understanding.

Videotapes can capture instruction in a manner which more closely resembles the form in which a teacher experiences it than any other assessment approach with the exception of classroom-based assessment approaches such as classroom observations and portfolios. There are important differences, however. On the one hand, a teacher has the luxury of not having to worry about classroom management problems and instructional problems simultaneously; on the other hand, the teacher is unable to test out ideas about possible problems or solutions by specific actions or questions. The most powerful example of the use of videotaped teaching episodes in evaluating teachers during the pilot test was a segment requiring teachers to interpret nonverbal cues from a student who was becoming increasingly discouraged by a series of negative comments on a paper he had written. Assessing a teacher's skills in using such nonverbal clues to diagnose problems in student motivation and/or understanding is one possible use of this assessment methodology. More generally, videotaped teaching episodes can moderately address the breadth of a teacher's instructional skills through the presentation of multiple vignettes which focus on a limited number of subjects/topics and types of students. Teachers can be asked about various aspects of instruction portrayed in the videotaped episodes, including communication, motivational methods, the way the teacher accommodates individual student differences (if these differences are made apparent to the viewer), and aspects of content pedagogy. If what teachers are asked to do focusses on higher-order thinking skills (e.g., compare what they have seen or to predict how the approach would work with another group of students), then the videotaped teaching episodes approach could provide a moderate amount of evidence as to the depth of a teacher's thinking with regard to instruction. However, videotapes do not capture the type of interactional skills which are a major part of most competencies within instruction, and hence reflect a very limited ability to represent skills in this domain.

Videotaped teaching episodes are probably the single best stimulus for measuring a teacher's ability to diagnose and evaluate student work which lends itself to representation through a videotape, such as a speech or dramatic presentation. If teachers provide a rationale for their evaluations, this assessment approach can provide extensive and authentic evidence as to the breadth and depth of a teacher's skills in evaluating this type of student work. Subject areas where this type of work is common include music, speech and drama, physical education, laboratory safety skills in science, and oral presentations in any subject. For other types of student work (e.g., written work), however, videotaped teaching episodes have a very limited ability to represent the breadth of skills in this domain. To measure a teacher's ability to evaluate these other types of student work, teachers might be asked to recognize different methods of monitoring and adjusting instruction during a lesson or to critique cne or two evaluation strategies portrayed. The ability of videotaped teaching episodes to assess the depth of skills in evaluating student learning again varies with the subject area, ranging from extensive for those subject areas where student performance lends itself to video representation, to hardly at all for other subject areas. It is difficult to imagine how skills in designing and interpreting written forms of evaluation, a major competency in this domain, could be efficiently assessed through a videotape stimulus. It is also difficult to imagine how to solicit a teacher response that would display much depth in thinking, as many of the skills in diagnosing and evaluating student learning depend on the teacher's knowledge of how to gather information which would not be represented in responses to a videotape of another teacher. Therefore, unless a breakthrough in this assessment technology occurs, it appears at this time that videotaped teaching episodes cannot evaluate the depth of a teacher's thinking in diagnosing and evaluating student learning in most subject areas. Similarly, it is difficult to see many ways in which the types of thinking represented in responding to videotaped teaching episodes would resemble those applied in actual teaching in most subject areas.

Participating in or creating a learning community does not readily lend itself to a visual representation, though a conference between a teacher and a specialist concerning a specific student could be portrayed and critiqued. Other skills, such as participating in a team working on school curricula, keeping abreast of developments in the subject area and/or grade level, and becoming familiar with and utilizing resources outside of the school, could also be portrayed, but it is difficult to devise a way in which teachers could respond to display their own skills in these areas. Therefore, videotaped teaching episodes seem to have virtually no ability to measure the breadth or depth of these skills for a teacher in any authentic way.

Subject areas vary in the degree to which they lend themselves to video portrayal. Video might be the best medium for conveying some aspects of some subject areas, such as science laboratory procedures, performance components in dance and physical education, presentations in speech and drama, and the presence or absence of safe conditions in science and physical education. In these instances, a teacher's response to a video stimulus can provide extensive evidence about both the depth of the teacher's subject-matter knowledge and skills in applying it. However, for many core subject areas, including English/language arts (excluding speech and drama), mathematics, social studies, and non-laboratory aspects of science, written materials are at least as good if not a better medium to portray subject matter. In these other subject areas, however, videotapes could portray alternative methods of representing or communicating concepts, excellent definitions and/or illustrations of concepts or processes, and egregious content errors. Thus the ability to portray the breadth of a teacher's subject-matter knowledge varies considerably with the subject area, but is moderate at best, limited to performance aspects, excellent communication of concepts and procedures, and egregious errors. However, while subject-matter knowledge is inherent in lessons, either explicitly or implicitly, the depth with which an observer can analyze a teacher's subject-matter knowledge is very limited. This would be true for teachers responding to videotaped lessons as well as for observers evaluating teachers through classroom observations, except in instances portraying student performance, where the depth of a teacher's subject-matter knowledge could be extensively evaluated. The authentic portrayal of subject-matter knowledge in many content areas is largely limited to a few aspects, such as recognition of errors and evaluations of definitions, metaphors, and explanations. However, if a teacher is applying their subject-matter knowledge to evaluation of a performance, then the authentic portrayal of subject-matter knowledge is extensive.

Multiple-choice Examinations

Multiple-choice examinations are presently the dominant form of assessment of teacher candidates, testing subject-matter knowledge through the various National Teacher Examinations and basic academic skills through numerous types of tests. This assessment approach has frequently been criticized by those who feel that most multiple-choice items and the thinking underlying responses to items do not resemble the thoughts and decisions in actual teaching. Since the multiple-choice examination assessment approach is familiar and well-documented, only a single instrument, which attempted to compensate for this weakness, was piloted. The instrument differed from other multiple-choice examinations primarily through the inclusion of "materials-based items" and the embedding of both theoretical and applied questions in classroom contexts. Materials-based items used documents which teachers are likely to encounter in their day-to-day teaching responsibilities, e.g, student worksheets, report cards, individual education plans. As an overview:

- The strength of the multiple-choice examination is in its ability to sample widely across those parts of teaching domains and domains which lend themselves to single, correct answers or a restricted number of reasonable answers. This is possible because of the large number of items which can be administered in a single examination.

- This strength in breadth is accomplished at the sacrifice of depth. While it is possible to design multiple-choice items which address higher-order thinking skills, the construction of equally attractive alternatives for such items is difficult.

- Because teaching is so complex that the correctness of a response depends on preceding and subsequent actions, as well as the specific context in which the response is given, the type of thinking teachers do is not reflected well in multiple-choice items; therefore, multiple-choice examinations are weak in authenticity. Furthermore, the relationship of performance on items to similar decisions in classrooms is unknown.

The materials-based items are an improvement over traditional multiple-choice items in terms of depth and application, but they are unable to overcome the constraints of the multiple-choice format. In addition, the materials-based items require more time for a response, limiting the number of items that can be included and reducing the major strength of the multiple-choice examination assessment approach.

In terms of planning and designing instruction, the materials-based items in the instrument piloted showed some potential to measure a teacher's skills in this teaching domain. Relevant items included sequencing a set of mathematics worksheets, selecting the most or least appropriate activity to teach a principle or concept, and demonstrating knowledge of appropriate instructional objectives and lesson plans. One could imagine more items of this type which address building on student resources and higher-order thinking skills. Items addressing more complex skills could be constructed, such as giving teachers a set of unit objectives and asking them to choose the most appropriate unit plan from a small group. However, only a limited number of such items can be included on a single multiple-choice examination, as the response time required for these items is much higher than for traditional items. Some skills in this domain would be extremely difficult to measure through multiple-choice items, such as the ability to evaluate a plan after implementation and to identify either the elements most responsible for its success or the most likely elements needing adaptation. Because of the difficulty in measuring complex skills in the planning and design of instruction, multiple-choice examinations can only moderately reflect the breadth of a teacher's knowledge of instruction. Some depth of knowledge in planning and instructional design can be indicated, in a very limited way, by asking the teacher to select alternatives identifying needed additions or problems in plans. However, it is impossible for single multiple-choice items to simultaneously capture many of the large number of factors that are considered by teachers in planning instruction. While some multiple-choice items piloted called upon a teacher to perform similar tasks to their everyday responsibilities (e.g., sequencing worksheets, choosing activities), teachers do much more in planning and designing instruction. Therefore, the authenticity of multiple-choice items is also very limited.

In the domain of classroom organization and management, multiple-choice items can cover virtually the complete range of skills, asking about principles of behavior management in a classroom context, the best choice of classroom organization for a specific objective, and the identification of activities which actively involve students and/or foster independent learning. The use of a video stimulus for multiple-choice items could address some skills in reading students' nonverbal cues. However, little depth of a teacher's knowledge is revealed, and the type of thinking bears virtually no resemblance to a teacher's thoughts and actions while they are rapidly processing information as they implement classroom and behavior management with actual students.

Multiple-choice items can moderately reflect skills in the area of instruction, such as knowing principles of subject-specific instruction, identifying more effective ways of motivating specific types of students, and knowing when and how to use strategies for accommodating individual differences in understanding and cultural and linguistic diversity within the classroom. However, skills such as tailoring strategies to individual students or classrooms, knowing when to apply a general principal and when an exception exists, and identifying and capitalizing on "teachable moments" and other interactional aspects of instruction cannot be measured well through multiple-choice items. Depth again is very limited by the multiple-choice response format, and the resemblance of the thinking required to respond to multiple-choice items in this domain bears little resemblance to the interactional thinking that is applied during the course of instruction.

In the domain of diagnosis and evaluation of student learning, multiple-choice items can address skills in diagnosing student errors and the most appropriate method of evaluating a specific objective. However, such items would be difficult to construct beyond a very superficial level for the skills of monitoring and adjusting instruction or for communicating student progress, so multiple-choice examinations can only moderately reflect the breadth of knowledge and skills in this domain. Although complex thinking is difficult to capture through multiple-choice items, materials-based items can reveal a limited degree of depth of teacher thinking. The decisions required can mimic some, but not many, of those needed in this domain as it is applied in and out of the classroom, so the authenticity of multiple-choice items in this domain is very limited.

In terms of participating in or creating a learning community, some items could address knowledge of current trends in subject-specific pedagogy and/or grade level psychology or instruction, and some materials-based items could assess a teacher's ability to identify student problems which would necessitate consulting resource specialists. However, items representing the full range of using school and non-school resources would be extremely difficult to construct, so the ability of multiple-choice items to measure breadth of knowledge in this domain is limited. Depth is again limited by the multiple-choice format, and the degree to which the items would reflect the ways that teachers actually participate in or create a learning community is virtually nil.

With respect to subject-matter knowledge, multiple-choice items can measure whether a teacher knows key concepts and applies them correctly, recognizes appropriate representations and/or concrete applications of concepts, and correctly applies standards of judging or valuing pertinent to the subject area. Multiple-choice tests are probably the single best assessment method for capturing the range of knowledge across the curriculum in a subject area, as the number of items included can be quite large for the time spent. Some depth can be captured, but mostly at the level of recognizing the correct answer rather than producing it. Therefore, multiple-choice examinations can extensively sample a teacher's subject-matter knowledge. With respect to depth of knowledge, however, the potential of multiple-choice examinations is very limited. Skills which involve higher-order thinking e.g., choosing correct representations of concepts such as division by fractions, are difficult to measure using multiple-choice items, as the construction of equally attractive but inaccurate alternatives is very challenging. It requires less depth of knowledge to recognize the correct answer, especially knowing that only one answer is correct, than to generate the correct answer. Multiple-choice questions can address some higher-order thinking skills and, at least in the case of mathematics, have been used to evaluate some depth of knowledge of concepts (i.e., recognition of an example of a concept -- Ball, 1990; McDiarmid et al., 1989; McDiarmid and Wilson, 1991). Materials-based items allow the measurement of a few aspects of application of subject-matter knowledge, such as sequencing worksheets and diagnosing student errors. However, teachers are often called upon to use their subject-matter knowledge to respond to student questions or to correct student misconceptions without the luxury of selecting from a menu of choices. Therefore, the ability of multiple-choice examinations to reflect ways in which teachers apply their subject-matter knowledge is also very limited.

Job-relatedness

All of the pilot tested assessment instruments had been developed in response to the criticism of the lack of job relevance on the part of many existing multiple-choice examinations and classroom observation instruments in use. Therefore, it is not surprising that the pilot tested prototypes contained elements of job relevance.

There are at least three ways in which an assessment approach can demonstrate job relevance:

- **Assessment materials**: The use of materials and tasks which were familiar to teachers, such as simulated student work or lesson plans
- **Performance-based emphasis**: The degree to which the assessment approach required the teacher to demonstrate a skill that was directly related to the classroom, e.g., evaluating student work or planning lessons
- **Predictive validity**: The degree to which performance on the assessment predicts performance on the job

The last dimension is the one that is of most interest to teachers and policymakers alike. Unfortunately, independent information about each teacher's on-the-job performance was not available to address this dimension for any of the assessment approaches. In fact, there are no instruments available to serve as measures of teaching performance, so investigation of this aspect of job relevance would involve costly multiple measures to establish the validity of an assessment approach.

Information was available to address the first two dimensions, which will be the focus of this section. Job-relatedness along one or both of these dimensions is a strength of most assessment approaches examined; summaries of the evaluation of each assessment approach along the two dimensions described appear in Table 2.

High-inference Classroom Observations

This assessment approach has no stimulus materials, so the first dimension is not applicable to this form of assessment. Classroom observations are the ultimate performance-based assessment, as teachers are observed actually teaching in their own classroom with their own students. The key to evaluating the job relevance of a specific

TABLE 2

JOB-RELATEDNESS

Assessment Approach	Degree of Resemblance of Stimulus Materials to Materials Encountered in Teaching	Degree to which Assessment Resembles Actual Teaching Experiences
High-inference Classroom Observations	N/A; no stimulus materials.	High, because teachers are observed teaching their students.
Portfolios	N/A; no stimulus materials.	High, because teachers are documenting their performance and student learning.
Semi-structured Interviews	High, if well-developed.	Moderate, if tasks include performance exercises.
Structured Simulation Tasks	High, if well developed.	Moderate, if tasks include performance exercises.
Performance-based Assessment Center Exercises	Variable, depending on the exercise.	Variable, depending on the exercise.
Videotaped Teaching Episodes	High, as teachers are portrayed instructing their students.	Not at all, as currently designed.
Multiple-choice Examinations	Moderate, if materials-based items are included.	Not at all, unless materials-based items are included in which case it is limited.

classroom observation instrument is assessing the degree to which the behaviors observed are indicative of specific skills which underlie good teaching. For a classroom observation instrument to be relevant to all teachers' jobs, it must be able to match behaviors in different teaching methods and widely varying contexts to more generally defined skills. To do this, some information about context and lesson objectives is needed for the observer to properly interpret the behaviors observed. The degree of match needed between the observer's experience and the teaching context is unclear, but grade-level experience, content knowledge, and possibly experience in similar contexts are potential areas of matching.

Portfolios

Like classroom observations, portfolios have no stimulus materials. To document the teaching of a unit, teachers collect handouts and corrected student work to supplement descriptions of the unit and lesson plans and reflective essays or logs. The handouts, comments on the student work, and often lesson and unit plans are the materials actually produced by the teacher in the course of teaching the unit. If classroom observations are the ultimate performance-based assessment of actual teaching, then portfolios come close behind.

Semi-structured Interviews

Piloted instruments representing this assessment approach use a variety of assessment materials commonly used in teaching, including textbooks, student work, illustrations, and passages of text. Well planned tasks can use stimulus materials such as these to ask teachers to perform tasks similar to those that they normally perform in the course of their teaching. In some piloted tasks, e.g., unit planning and historical interpretation, teachers were asked to do tasks which beginning teachers probably do not do routinely, but which expert teachers in the subject matter agree that teachers need to perform in order to teach the subject matter effectively. Sometimes the tasks may be in areas in which beginning teachers are not highly skilled, e.g., evaluating student learning, but some demonstration of progress toward mastery may be expected. These tasks only bear a moderate resemblance to those that the teacher performs in the classroom, however. A classroom teacher has much more information available about student backgrounds and interests and previous learning, which should affect the way they approach their teaching tasks.

Structured Simulation Tasks

Structured simulation tasks can be designed so that teachers are asked to perform tasks that are highly reflective of teaching duties. Like semi-structured interviews, piloted structured simulation tasks used a variety of materials that teachers use or produce as they teach: Lesson and unit plans, letters from parents, district memos, and student work. As with other assessment approaches simulating performance, the information provided in the simulation cannot duplicate the amount that a teacher might consider when actually performing tasks, so the resemblance to actual performance of tasks can only be moderate.

Performance-based Assessment Center Exercises

This approach to assessment typically focusses on the performance of tasks in order to demonstrate specific skills, and thus it has similar potential with regard to job relevance as semi-structured interviews and structured simulation tasks. The assessment materials used vary with the specific exercise, but if the focal skills are job-relevant, then it is likely that materials can be used which resemble those which teachers normally use. The key to evaluating the job relevance of particular performance-based assessment center exercises is to consider the relevance of the skills being assessed to those used in teaching the subject matter.

Videotaped Teaching Episodes

The stimulus for this form of assessment is a series of videotaped portrayals of teachers instructing their own students, a very familiar experience for teachers. Although the scoring system used for the piloted prototype did not focus on teacher performance in any depth, analysis of the prototype suggested modifications which might be useful in eliciting teacher performance. Segments portraying skills in evaluating and responding to verbal and nonverbal student cues are difficult to simulate by other assessment approaches; videotape seems the ideal medium for soliciting evidence about these skills. However, the piloted scoring system focussed on questions designed to be answered in brief responses. This did not begin to capture a teacher's thinking as it occurs during instruction or during reflection on previous instruction. More elaborated responses, perhaps scored holistically, seem to be a more appropriate scoring format for this assessment approach. This would increase the resemblance of teacher responses to types of thinking which occur during teaching and reflection on teaching though, as with other simulations, the resemblance is likely to be moderate, at best.

Multiple-choice Examinations

This type of assessment has been criticized for its lack of job relevance. The pilot tested instrument addressed this criticism through the increased use of familiar materials. Such items required teachers to choose activities, sequence instructional materials, and analyze simulated materials (e.g., student work, lesson plans, report cards, sets of work sheets) which teachers encounter in their daily work. The inclusion of these "materials-based items" increased the resemblance of the stimulus materials to those that teachers encounter in their daily experience. However, these items consume much more time than more conventional multiple-choice items, placing limits on either the number of such items included or the total number of items on the test. Even with the inclusion of this type of item, the multiple-choice format provides only a limited amount of information about a teacher's ability to perform these tasks in a classroom.

Appropriateness for Beginning Teachers

The appropriateness of an approach for assessing beginning teachers is strongly affected by the content of the instruments which represent that approach. Since content varies from instrument to instrument within an approach, and only a small sample of instruments were piloted that represent each approach, it is difficult to draw conclusions about the appropriateness of an approach as a whole. Therefore, the discussion is focussed on the appropriateness of each pilot tested instrument, providing tentative conclusions, where appropriate, about the assessment approach represented.

Before discussing each instrument, however, there are some generalizations about the assessment of beginning teachers that appear to hold across assessment instruments. Full-time teaching experience seems to make an important contribution to the development of a teacher's skills. Many second-year teachers participating in different assessments commented that they would not have done as well the first year. This was not systematically investigated, but the nature of the specific teacher comments suggest that there are opportunities for teachers to learn from experience which are more likely to occur in the second year of teaching than in the first:

- First-year teachers do not have much time to reflect on their teaching. They generally have few or no previously developed lesson plans, and must develop or adapt most lessons they teach. In addition, they are learning about their particular students and the school and district in which they teach. In the second year, if their teaching assignment is the same, there is more time for reflection, as they can use lessons previously designed.

- Teaching a lesson for the second time enables a teacher to make comparisons with previous instruction and to begin to distinguish between problems particular to a group of students and problems which seem to be inherent to a lesson.

Beginning teachers showed a pattern of weaknesses across assessment instruments and approaches, which were not always consistent with those mentioned in the literature on beginning teachers. For example, although classroom management is almost universally listed as a problem for beginning teachers (Veenman, 1984), this was not a common area of weakness detected by the classroom observation instruments piloted. Weaknesses observed included:

- **Content knowledge.** Weaknesses in content knowledge were observed in virtually every pilot test. These content weaknesses appeared across multiple subjects. Teachers had difficulty in sequencing instruction, accurately constructing explanations or representations (e.g., analogies, models) of concepts and principles, evaluating student work, and citing real-life applications of concepts or principles. These content weaknesses will severely impact a teacher's ability to implement the State Curriculum Frameworks, and do not seem to be confined to beginning teachers in California, as they are consistent with studies conducted by the National Center for Research on Teacher Education (McDiarmid et al., 1989; McDiarmid and Wilson, 1991.)

- **Subject-specific pedagogy.** The weaknesses in content knowledge mentioned above exacerbate weaknesses in subject-specific pedagogical skills. Although the pilot-test teachers found designing lesson plans one of the easier assessment tasks, performances on a variety of assessment instruments suggest that many teachers have difficulty in estimating the amount of time various activities take, in anticipating student problems, and in structuring activities to increase the amount of higher-order thinking and/or depth of thinking of students. A reflective teacher's skill in estimating time requirements and in anticipating student problems is likely to increase with experience. However, learning how to better structure activities and how to extend learning in successful activities is more likely to require outside support.

- **Teaching diverse students**. Many beginning teachers participating in the pilot tests exhibited a keen knowledge of their students, and were able to design activities in semi-structured interviews and portfolios which exhibited knowledge of the backgrounds, interests, and dislikes of their students. This knowledge of students did not extend, however, to students with whom they were unfamiliar. Since research shows that knowledge of students is a complex skill more characteristic of experienced teachers than beginning teachers (Leinhardt, 1983), it is possible that only years of teaching, which can provide experience with an increasing variety of students, provides a sufficient knowledge base upon which teachers can draw to make inferences about a wide range of students.

The pilot tests were not designed to systematically explore the knowledge, skills, and abilities of beginning teachers, but ensuring that assessments are appropriate for beginning teachers requires some knowledge of beginning teacher skills. If one goal of teacher assessment is to improve the quality of teaching in California schools, it is especially important to understand which skills may be expected to be relatively developed and which may appear only in embryonic form. It is possible that these understandings may shift over time with changes in teacher preparation. Some critics of education in the United States have turned their attention to teacher preparation and identified areas for improvement (Goodlad, 1990; Ball and Wilson, 1990; Leinhardt, 1988; McDiarmid, et.al., 1989). However, few of the proposals have been implemented for a sufficient period to distinguish between areas which could be strengthened with redesigned preparation and those which require additional teaching experience. Given that current research is only suggestive, assessment designers must be prepared to monitor the emerging literature to remain informed of significant findings concerning beginning teacher skills and adapt or perhaps even reorient the assessment to measure different skills.

The appropriateness of each pilot tested instrument for assessing beginning teachers is summarized in Table 3, which presents three pieces of information about each instrument as potential indicators of appropriateness for beginning teachers. The first two indicators are statistics from the pilot testing experience: The percentage and number of teachers who agreed that they had adequate preparation to perform the assessment tasks, and the general level of performance. To assist the reader in interpreting these statistics, limitations of the data are discussed here. Although teachers sometimes received at most a general description of how the various tasks were to be scored, they never saw their own individual scores, and often didn't even see the scoring protocols. Their estimation of the difficulty of the assessment, as indicated by which tasks they considered to be easy and which to be hard, did not coincide with their scores. For instance, teachers invariably saw lesson

TABLE 3

APPROPRIATENESS FOR BEGINNING TEACHERS

Assessment Approach and Instrument	Percent and Number of Sample Teachers who Rated their Preparation for the Assessment as Adequate	General Level of Performance of Sample Teachers	Diagnostic Capability
Classroom Observations			
CCI	56 %* (10 of 18)	80 % (33 of the 41) teachers received "acceptable" ratings on at least seven of the ten indicators.	Depends on nature of feedback. Teachers are rated on ten dimensions of teaching. If reasons for failing are communicated, diagnostic capability for the lesson observed is high. Generalizability to other lessons, especially to quite different topics and subject areas, is likely to be poor; the extent to which multiple observations solves this problem is unclear.
Science Laboratory Assessment	72 % (21 of 29)	The percentage of teachers passing each domain ranged from 84 % (Content) to 100 % (Communication), but passing standards were not clearly set.	If the present scoring system is refined; potential may be similar to that of the CCI.

*Not all teachers completed the questionnarie.

TABLE 3 (cont'd)

APPROPRIATENESS FOR BEGINNING TEACHERS

Assessment Approach and Instrument	Percent and Number of Sample Teachers who Rated their Preparation for the Assessment as Adequate	General Level of Performance of Sample Teachers	Diagnostic Capability
Portfolios			
	93% (14 of 15)	The percentage of teachers clearly passing each evaluation category ranged from 37% (unit design, general pedagogical abilities, subject-specific pedagogical abilities) to 56% (reflective ability).	Depends on the nature of feedback. If detailed feedback is provided, diagnostic capability for the unit taught is high. Generalizability to other topics is unknown.
Semi-structured Interviews			
SSI-SM	80% (16 to 20)	The average rating on the two tasks scored was "marginal", but this was partially an artifact of weaknesses in the instrument.	High with respect to the specific topics covered; generalizability is unknown.
SSI-EM	70% (31 or 40)	Percentage of teachers passing ranged from 22 % on Topic Sequencing (Fractions) to 63 % on Instructional Vignettes.	Present scoring system needs reconceptualization before diagnostic capability can be estimated.

TABLE 3 (cont'd)

APPROPRIATENESS FOR BEGINNING TEACHERS

Assessment Approach and Instrument	Percent and Number of Sample Teachers who Rated their Preparation to do the Instrument Tasks as Adequate	General Level of Performance of Sample Teachers	Diagnostic Capability
SSI-SSS	94% (15 of 16)	Not yet scored.	Potential is expected to parallel that of SSI-SM.
Structured Simulation Tasks			
Secondary Life/General Science Teacher Assessment	76%* (35 of 46)	Teacher scores on the tasks ranged from a low of 16 % of the total possible points (for critiquing a specific lesson) to a high of 61 % of total possible points (for writing a letter to parents and students explaining why a certain course was important and of value to students.)	Poor, because there is not enough information on any one skill to allow generalization. This test was not constructed to give diagnostic information.

TABLE 3 (cont'd)

APPROPRIATENESS FOR BEGINNING

Assessment Approach and Instrument	Percent and Number of Sample Teachers who Rated their Preparation to do the Instrument Tasks as Adequate	General Level of Performance of Sample Teachers	Diagnostic Capability
Secondary English Teacher Assessment	90% (of)		Poor, because there is not enough information on any one skill to allow generalization. This test was not constructed to give diagnostic information.
Assessment of Competence in Monitoring Student Achievement in the Classroom	57% (24 of 42)	The mean performance of groups of teachers on the two forms ranged from 38 % of the total possible points for pretest teachers participating in staff development to 47 % of the total possible points for posttest teachers not participating in staff development.	Poor, because not enough information on any one skill to allow generalization.
Performance-based Assessment Center Exercises			
Secondary English Assessment	63% (18 of 19)	Varied according to task, ranging from 63% to 95% received passing rating from both raters.	High, assuming the scorers were trained to rate subscales.

TABLE 3 (cont'd)

APPROPRIATENESS FOR BEGINNING TEACHERS

Assessment Approach and Instrument	Percent and Number of Sample Teachers who Rated their Preparation to do the Instrument Tasks as Adequate	General Level of Performance of Sample Teachers	Diagnostic Capability
Videotaped Teaching Episodes			
LAPKA	78% (33 of 42)	Varied by scenario, with means ranging from 57 % correct to 87 %.	Poor, because not enough information on any one skill to allow generalization.
Multiple-choice Examinations			
Elementary Education Examination	83% (114 of 137)	Average scores ranged from 68% in mathematicis and social studies to 74% in language arts.	Poor. This test was not constructed to give diagnostic information with respect to specific skills. Subject matter subscores were available, but the items differed greatly between subjects.

planning as one of the easier tasks, presumably because of the high frequency with which they do the task. However, as a group, they also consistently received low scores on lesson planning tasks for every assessment approach piloted. It is possible that the teachers' assessments of their preparation are probably more reflective of their familiarity with these tasks from teacher preparation or from teaching than of their level of proficiency.

The way that the performance data is summarized for each instrument varies with the type of scoring system. Some systems did not attempt to set passing scores, but only yielded numerical scores. Other scoring systems concentrated on setting pass/fail standards for individual tasks, but not for the collective set of tasks. When a passing standard is available, data is presented in the form of the range of the percentage of teachers passing individual tasks. When no passing standard is available, the range of mean scores of teachers across individual tasks is presented. These mean scores are presented in the form of the percentage of total points possible. The major limitation of these data is that sometimes teacher scores were somewhat depressed because of flaws in the stimulus or scoring materials.

The third indicator of the appropriateness of an assessment instrument for beginning teachers is a summary of FWL staff evaluations of its diagnostic capability, operationally defined as the quality of information available to guide staff development.

High-inference Classroom Observations

Table 3 begins with the two examples of the classroom observations assessment approach that were piloted: The CCI and the Science Laboratory Assessment. The CCI focussed on a broad variety of general pedagogical skills; the Science Laboratory Assessment focussed on both general pedagogical skills and on subject-specific skills related to science laboratory instruction. Although only a little over half of the teachers participating in the pilot test of the CCI believed that they had had adequate preparation to perform the tasks, over 80% received passing scores on at least seven of the ten indicators, which was suggested as the passing criterion. A higher percentage passed each of the seven domains of the Science Laboratory Assessment. However, passing levels for the latter instrument were not clearly defined, leaving much up to the judgment of the individual observers, who at least occasionally applied different standards.

Classroom observations potentially provide a rich source of diagnostic information and have traditionally been the primary diagnostic method used by districts. Information available from these high-inference classroom observations includes both a summative assessment (pass/fail) for each indicator and summary notes made by the observer which generally describe patterns of behaviors leading to the pass/fail decision. Based on this information, teachers failing a particular indicator could be provided assistance targeted to the specific indicator. Staff development sessions could orient the teacher to a fuller understanding of the importance of the indicator and various methods of achieving it. More substantial formative feedback on a teacher's efforts to improve their performance would depend on information gathered through a series of classroom observations. This type of assistance is sometimes provided by mentor teachers. However, mentor teachers often do not have the time to become familiar enough with recurrent patterns of behavior of a beginning teacher in the depth needed to provided assistance directed to a specific competency. Moreover, data at the level of detail contained in observation summaries, and not just pass/fail data according to indicator, are needed to identify strengths and weaknesses observed and to suggest the areas of assistance needed by a beginning teacher.

Another diagnostic problem with classroom observations is that they provide unstable descriptions of behavior, because patterns of behavior vary from one observation to another. Individual teachers also tend to exhibit patterns of behavior that vary with the subject matter taught (Stodolsky, 1988). Multiple observations are used to compensate for this difficulty, but it is not clear whether multiple observations are able to compensate for differences between subject matter or lesson objectives. High-inference classroom observation instruments use professional judgments based on the appropriateness, rather than the frequency, of specific behaviors. This feature can probably compensate for behavior differences inherent in different topics or lesson objectives. However, it is likely that teachers are not equally skilled across different topics and instructional activities. Reconciling different results from different observations might be difficult, especially if different lesson topics, student groupings, or lesson objectives occur. However, this sampling and generalizability problem is not unique to classroom observations, but occurs for all other assessment approaches as well.

Portfolios

Nearly every teacher participating in the portfolio assessment believed that they had adequate preparation to produce the portfolio. However, the percentage of teachers clearly passing each category (i.e., receiving a passing score from each of the two raters) was disappointingly small. The beginning teachers identified long-range planning and planning for a unit with which they had had little or no experience as areas of difficulty. Though some of the individual lessons were exemplary, many of the teachers had trouble identifying and sequencing activities to facilitate the accomplishment of identified goals; these difficulties in planning and sequencing instruction were also found in teachers participating in other assessment approaches.

The diagnostic capability of the portfolio method of assessment depends on the nature of the feedback provided. Detailed comments on the portfolio identifying strengths and weaknesses in specific skills with specific examples would provide a high degree of feedback with respect to the particular lesson. The generalizability to other lessons, however, is unknown.

Semi-structured Interviews

Most teacher participants believed that they had been adequately prepared to perform the tasks comprising each semi-structured interview assessment. However, scores were not high. In the case of the SSI-SM, this was due at least partially to a mismatch between the questions asked and the scoring system, which was developed after the instrument was administered. Subsequent revisions to produce a closer match have improved the scoring system, which produces a rich amount of information about the depth and breadth with which a teacher is able to talk about teaching a specific topic. What is still unclear, however, is the relationship between a teacher's ability to describe what they might do with what they actually do in the classroom for the topics which they do teach.

The scoring system of the SSI-EM varied substantially by task and by topic within task, and was not deemed suitable for statewide administration due to anticipated problems with respect to reliability, validity, and fairness. The scoring methodology used for the SSI-SM and projected for use with the SSI-SSS is holistically based, and focussed on three major teaching domains. An Evidence Summary document summarizes evidence from the interview which supports a rating for each competency within a domain. This document, if provided to a beginning teacher and/or support provider, could provide detailed diagnostic information with respect to the competencies rated on the particular topic examined.

However, the subject which was the focus of two interviews, mathematics, is hierarchically and sequentially structured so that "basic" topics can be chosen for the focus of the interviews. It can be argued, therefore, that the performance ought to generalize to other topics. Other subjects (e.g., science, social science, English) are not organized hierarchically. It is not clear that performances focussing on one topic in these subject areas are generalizable to other topics.

Structured Simulation Tasks

While most of the secondary teachers participating in the Secondary Life/General Science Teacher Assessment believed that they had been adequately prepared, only a slight majority of the elementary teachers participating in the Assessment of Competence in Monitoring Student Achievement in the Classroom felt adequately prepared. Compared to the total number of points possible, however, neither group of teachers did particularly well, on the average. While the scoring methodology (i.e., a teacher's response is compared to a set of previously determined answers) facilitates scoring, the performance is so dependent on simultaneous knowledge of content, type of student, and instructional activity that the generalizability may be limited.

The science teachers completing the Secondary Life/General Science Teacher Assessment complained that they were not familiar with some of the topics and instructional practices portrayed in the assessment. Simultaneously considering the topic, type of student, and instructional activity specified in a task may be too complex for beginning teachers, whose experience with the specific content and context in a task is at best limited and may be nonexistent. The less familiar teachers are with the topic, type of student, and instructional activity portrayed in the stimulus materials, the more difficult it will be for them to display the knowledge and skills they do possess.

These assessments were designed to produce a licensure decision in a cost-effective manner, and not to provide diagnostic information. The assessment methodology depends on sampling across tasks, types of students, topics, and instructional approaches to reach a decision about a teacher's basic teaching competence in a specific subject. Little specific information about particular competencies is provided, and that information is affected by the degree of a teacher's familiarity with the topic, type of student, and teaching approach. Experience with a parallel assessment for lawyers suggests that correlations of performance across families of tasks (e.g., lesson planning vs. evaluating student work) are roughly the same as correlations of performance within a family of tasks when one or more elements are changed. Therefore, it is impossible to generate a score for a particular competency (e.g., lesson planning), and this assessment methodology has poor diagnostic ability.

Performance-based Assessment Center Exercises

Roughly two-thirds of the teachers believed that they had been adequately prepared for the assessment. Not surprisingly, since the tasks focussed on distinct and independent skills, the average performance varied among the assessment tasks. The highest percentage of teachers passing occurred for the discussion of a literary work, and the lowest occurred for the activity which required teachers to respond to student writing samples. (The passing percentage for the latter, however, was greatly diminished by a large number of teachers whom one scorer was unable to rate.)

Since the scoring of each task includes a number of subcategories, in-depth diagnostic information could be generated for each specific skill being measured by a task, provided that scorers are trained to rate the subscales. It was not clear whether the subcategories should be used only to guide the general overall rating or should be rated independently. If the former approach is used, and no other feedback is provided to the teacher, then the exercises will be of limited diagnostic use.

Videotaped Teaching Episodes

Again, most of the teachers believed that they were well-prepared for the assessment. Performance, however, varied by scenario, with averages ranging from 57% of the total possible points for the scenario depicting writing assessment to 87% for one of the scenarios depicting large group reading instruction. (Average scores for the other two scenarios were 68% and 69%.) A scoring methodology similar to that of the structured simulation tasks was used (i.e., comparing teacher responses to previously determined answers.)

The diagnostic ability of the scoring system of the piloted prototype was poor. Teachers provided short-answer responses to a variety of questions. While a score for a specific method of language arts instruction can be calculated, the scoring methodology cannot indicate further diagnostic information about why a teacher scores well or poorly on a specific scenario. Furthermore, scorers questioned the validity of the scoring system, noting that some teachers provided correct short-answer responses, then elaborated their answer and contradicted themselves or made it plain that they did not understand the short response that they had provided. As mentioned elsewhere in this report, a prototype that takes full advantage of the videotape stimulus has yet to be developed, so no conclusions can be drawn about the diagnostic potential of videotaped teaching episodes as an assessment methodology.

Multiple-choice Examinations

This assessment approach was represented by the Elementary Education Examination with its materials-based items. Most teachers believed that they had been adequately prepared for the examination, and the average scores indicated that the teachers correctly answered between two-thirds and three-fourths of the items. While subscores were computed by subject matter, the items varied substantially between subjects. For example, the main focus of the language arts items was on correctly characterizing methods of teaching reading, while the main focus of science items was on content knowledge. Items measuring pedagogical content knowledge proved to be especially difficult to construct. Although one could imagine a multiple-choice examination focussing on one particular teaching skill, the strength of the multiple-choice examination is its breadth, not depth (Estes et al., 1990), and many of the skills desired in teachers appear to be extremely difficult to measure using multiple-choice items. Therefore, the diagnostic capability of multiple-choice examinations could be expected to be high in terms of breadth of skills, and poor in terms of depth.

Appropriateness across Different Teaching Contexts

Before discussing the appropriateness of individual assessment approaches across teaching contexts, two problems which apply across assessment approaches will be described:

- Assessing instruction in unfamiliar contexts

- Focus of instruments on the upper grades covered by a credential

Assessing Instruction in Unfamiliar Contexts

In the previous section, the difficulties that beginning teachers have in describing and analyzing instruction with which they have little or no experience were described. Teacher responses and criticism of the piloted assessment instruments indicate that their responses were almost always based on their experience with their own students, even when they were explicitly directed to focus on another type of student. Generalizing to different types of students and communities requires not only theoretical knowledge of important factors which differentiate students and communities but also some experiential knowledge of how these factors interrelate and affect instructional decisions. More depth of knowledge can be expected of teachers when it is related to the contexts in which they have taught.

For credentialing purposes, a teacher assessment must sample across grade levels, topics, and types of students covered by the credential. However, beginning teachers are likely to be able to demonstrate in-depth knowledge for only the limited number of instructional techniques, grade levels, topics, or types of students with which they have had experience. When the topic, the type of students, and the teaching approach are simultaneously specified in the assessment materials to which the teacher is asked to respond, the odds increase that a teacher has had little or no experience with one or more of the areas in the specified combination. In assessment approaches using this methodology, a teacher's lack of knowledge in one area may affect the ability of the assessment to reflect their knowledge in other areas. While the assessment evaluation in these cases may accurately represent the skills in applying that knowledge with a particular group of students, it is likely to underrepresent a teacher's knowledge in specific areas. After one or two years of experience, beginning teachers have some experience with the appropriate pacing of instruction, appropriate level of complexity of materials, and difficulty of specific concepts for the types of students that they teach. When asked to plan for a different type of student, the beginning teacher is more likely to make errors in generalizing from both their theoretical knowledge and their experience. The less the teaching context resembles that with which the teacher has had experience, the more egregious the errors are likely to be.

In addition, our consultant on cultural diversity, Dr. Sharon Nelson-Barber of Stanford University, worried that some teachers might be penalized for answers based on their experience with effective instruction in their own setting, but which are labelled incorrect by the assessment. For example, the current trend in teaching writing is to deemphasize mechanics (e.g., grammar and punctuation) and to concentrate on content; however, many inner city teachers and parents feel strongly that skills in the mechanics of writing are especially critical for their students. Such a teacher, when asked to respond to student writing in an assessment exercise, is likely to correct spelling, grammar, and punctuation as well as to comment on the substance of the student's writing. However, the correction of mechanical errors was penalized in the scoring systems of several of the instruments piloted, on the grounds that an emphasis on mechanics inhibits the writing process. This example is an instance where there are differing opinions on what constitutes an appropriate teaching strategy. Beginning teachers may also be penalized when they overgeneralize from their limited experience to suggest inappropriate teaching strategies for a particular group of students due to their unfamiliarity with that group of students.

Assessment approaches whose scoring methodology consists of the comparison of responses to previously determined correct ones (e.g., structured simulation tasks, multiple-choice examinations and the instrument representing videotaped teaching episodes) would be most difficult to adapt in order to address this concern, as the teacher does not provide a rationale for their decisions. More open-ended scoring systems, e.g., the semi-structured interview and holistically scored exercises in the performance-based assessment center exercises, have more potential to distinguish between beginning teachers who are overgeneralizing from their limited, but well-grounded, experience and beginning teachers whose ideas are relatively uninformed by either educational theory or experience. The potential of these open-ended scoring systems to avoid penalizing beginning teachers for their lack of experience depends on: (1) the depth of the rationale offered by the teacher, (2) the identification of the type of students with whom the teacher has had experience, and (3) the training, and possibly experience, of the scorers.

Because of the short period of time allotted to teacher preparation in California, it is likely that a teacher will teach, or observe others teaching, only a limited range of students. It is only these types of students about whom a teacher can be expected to exhibit a depth of knowledge. If State policy requires familiarity with specific groups of students, then student teaching should occur in contexts where these students are represented. There is a limit, however, to the knowledge and skills which a teacher may be expected to acquire during student teaching. At some point, increasing the number of contexts in which student teaching occurs may only increase a student teacher's confusion in trying to process incoming information rather than increase the breadth of knowledge, and may also decrease the depth of knowledge acquired. This presents a dilemma for beginning teacher assessment, suggesting that broad sampling of teaching contexts can only be accomplished through limiting expectations for the depth of responses.

Classroom-based assessment approaches (i.e., classroom observations, portfolios) have extremely limited sampling ability. Simulation approaches have more ability to sample across contexts, but unless a teacher is explicitly asked to compare instructional approaches in two or more contexts, the measurement of ability to teach in a specific context is confounded with the measurement of specific skills or abilities which are the focus of the task or exercise. Multiple-choice examinations were the only assessment approach with an ability to broadly sample across varying teaching contexts, but this was at the level of knowledge only, not application, and at the expense of the depth of knowledge assessed. The remaining assessment approaches showed very limited ability to measure a teacher's ability to teach in varying teaching contexts. While it would be ideal to only credential teachers who can display the ability to teach in various teaching contexts, it is perhaps unfeasible to expect this of beginning teachers. First- and second-year teachers are only beginning to master their craft in a particular context and typically do not have the breadth of experience to enable confident generalization from their experience. Therefore, any skills in teaching in contexts with which they ar not familiar are likely to be weak.

Focus of Assessments on the Upper Grades Covered by a Credential

With one exception, LAPKA (the videotaped teaching episodes instrument), the piloted assessment instruments which were not classroom-based did not sample across the range of grades covered by a credential, but typically concentrated on the upper grades: High school for single subject credentials, and the intermediate grades (grades 4-6) for the multiple subjects credential. Assessment of subject-specific pedagogical skills in the primary grades (grades K-2) was rare. The lists of beginning teachers obtained from districts both within and outside the CNTP suggests that beginning secondary teachers are most likely to teach at the junior high school or middle school levels; beginning elementary school teachers seem to be more evenly distributed across grades. In order to be fair, if the policy is to measure the breadth of a teacher's skills across the grades represented by the credential, grades covered by the credential should be equally represented.

It is likely that this focus on upper grades is not entirely due to the tastes of the assessment developers. It seems to be more difficult to assess how teachers simplify concepts and principles for purposes of introduction than to assess how they elaborate them in some depth. It seems especially problematic to evaluate how a primary-grade teacher instructs in subjects such as science and social science where basic skills are not as well defined and researched as in language arts and mathematics.

The above discussion illustrates issues to be discussed in setting policy for beginning teacher assessment. Because the issue of breadth vs. depth of knowledge of teaching contexts has not yet been settled, Table 4 uses two factors to indicate the ability of an assessment approach to measure appropriateness across teaching contexts. The first factor summarizes conclusions about the ability of each assessment approach to measure a teacher's applied skills in teaching in diverse settings. The second factor is an indication of an assessment approach's ability to measure depth of knowledge: The ability to accommodate different conceptions of teaching, i.e., the flexibility to allow teachers to display the skills which they believe they have mastered.

High-inference Classroom Observations

In terms of sampling ability across teaching contexts, observations are limited by the teaching responsibilities of the teacher to be observed. Elementary teachers almost always teach a single classroom of students in a single teaching context. Because they teach multiple subjects, elementary teachers can be observed teaching different subjects, teaching different topics within a single subject area, or using different instructional groupings.

TABLE 4

APPROPRIATENESS ACROSS DIFFERENT TEACHING CONTEXTS

Assessment Approach	Ability to Measure a Teacher's Skills in Teaching in Diverse Settings	Ability to Accommodate Different Conceptions of Teaching
High-inference Classroom Observations	Varies with the extent of diversity among the students taught by an individual teacher, the extent of information about students available to the observer, and the training and experience of the observer.	High potential, depending on training, and perhaps experience, of observer, as well as the comprehensiveness of the definition of the skills which are assessed.
Portfolios	Limited by the variation in students and the teaching responsibilities of the teacher being assessed. Documentation of more than one class would likely increase the burden on a teacher or reduce the depth of information available.	High, depending on the training, and perhaps experience, of the observer, as well as the comprehensiveness of the definition of the skills or knowledge which are assessed.
Semi-structured Interviews	Limited, as only a few types of students and/or teaching techniques can be specified, and the relationship of assessment performance to application is unknown. Some potential for separately measuring knowledge with respect to type of student, teaching technique, and/or topic.	High potential, depending on training, and perhaps experience, of scorers, as well as on the definition of skills to be rated.

TABLE 4 (cont'd)

APPROPRIATENESS ACROSS DIFFERENT TEACHING CONTEXTS

Assessment Approach	Ability to Measure a Teacher's Skills in Teaching in Diverse Settings	Ability to Accommodate Different Conceptions of Teaching
Structured Simulation Tasks	Limited, as only a few combinations of students, topics and/or teaching techniques can be presented in tasks, and knowledge of each of these areas cannot be measured separately. High potential for measuring application of skills in specific situations where a consensus exists about appropriate instruction.	Limited, as scoring criteria are predetermined and are unlikely to encompass radically different conceptions of teaching.
Performance-based Assessment Center Exercises	Limited, as only a few combinations of students, topics and/or teaching techniques can be presented.	Variable, depending on the nature of tasks and training of scorers.
Videotaped Teaching Episodes	A limited number of combinations can be portrayed. Potentially high, in terms of measuring recognition of applied principles of effective instruction of specific types of students.	Scoring system needs reconceptualization before ability can be assessed.

TABLE 4 (cont'd)

APPROPRIATENESS ACROSS DIFFERENT TEACHING CONTEXTS

Assessment Approach	Ability to Measure a Teacher's Skills in Teaching in Diverse Settings	Ability to Accommodate Different Conceptions for Teaching
Multiple-choice Examinations	High, in terms of sampling knowledge of generally accepted principles. Extremely limited, in terms of application of skills.	High, in terms of assessing knowledge of a number of teaching techniques. Limited, with respect to assessing skills in methods commonly used by a teacher, as correct answers neither take into account the context of a teacher's experience nor necessarily focus on teaching techniques with which the teacher is familiar.

However, the type of student cannot be varied beyond the natural variation in the teacher's classroom. Secondary and middle school teachers typically teach multiple classes of students. To the extent that students in these classes vary (e.g., achievement level, cultural background, linguistic facility with English), then a teacher's ability to teach different types of students can be observed through repeated observations, each of which focus on a different class of students. Thus the sampling ability of high-inference classroom observations with respect to context depends on the variance in students taught by each individual teacher.

The ability of high-inference classroom observations to measure a teacher's skills in teaching in diverse settings varies not only with the extent of variation among students taught by an individual teacher. It also is dependent on the extent of information about students which is available to the observer, the quality of training of observers, and the extent of the specific observer's experience with the type(s) of students in the classroom observed. Variation among students is not always apparent, however, and an observer is frequently dependent upon information about students provided by the teacher observed. Moreover, teachers, and especially beginning teachers, vary in their knowledge of their students, especially in the familiarity with their students' cultural backgrounds and the ability to identify a specific student's learning problems. An observer spends approximately one hour in the classroom -- enough time to identify gross discrepancies between the information received from the teacher and events observed, but not enough time to verify the teacher's description or to judge the appropriateness of the instruction of specific students.

The high-inference classroom observation instruments piloted were used to observe a wide variety of instructional approaches, ranging from whole class instruction to the management of work in small groups. It appeared that the piloted instruments were adaptable to every teaching situation with the possible exception of a vocational education class. This class centered around students' repetitive practice of previously learned skills where little instruction is evident. Thus, the high-inference classroom observation assessment approach seems to have high potential to accommodate different conceptions of teaching, providing teachers with an opportunity to demonstrate what they consider their strengths. To do this well, however, the competencies on which a teacher is rated must be defined so that they clearly apply across different teaching approaches and classroom contexts. In addition, observers must be trained to recognize the competencies in varying teaching contexts and varying teaching approaches. A match between the observer's experience and the grade level, subject matter, and teaching context of the teacher being observed may facilitate the ability of the observer to accurately evaluate the teacher.

Portfolios

Like high-inference classroom observations, portfolios also utilize a teacher's experience teaching his/her own students. The sampling potential of portfolios across teaching contexts is the same as classroom observations, i.e., it depends on the extent of variation in a teacher's classroom and teaching responsibilities. Because of this restricted sampling ability, portfolios are extremely limited in their ability to evaluate a teacher's ability to teach in diverse settings.

It would be possible to ask a teacher to document the teaching in more than one class. To do so would vastly increase the amount of documentation, if the same number of documents, logs, and reflective essays were required for each class. If a teacher is asked to document the teaching of various types of students, drawing examples from different classes, this approach would weaken the portfolio approach's ability to evaluate the teaching because evaluators use different segments of a portfolio to provide cumulative and corroborating evidence of the appropriateness of a teacher's instructional decisions. This type of evidence would not be available if the documentation was drawn from different classes.

Since portfolios can allow the teacher to choose the unit documented, teachers have an opportunity, within the constraints of their teaching assignments, to choose the classroom, topic and teaching approaches to demonstrate their strengths in instruction in that context. However, as with classroom observations, valid and reliable assessment of a teacher's knowledge and skills through a portfolio depends on two factors. The first is defining the competencies assessed so that they are specific enough to be clear yet apply across various teaching contexts. The other factor is the training of the scorers so that they can recognize evidence pertaining to each competency regardless of the teaching approach, topic, or type of students taught. Matching the experience of the scorer to the teaching context of the teacher might facilitate scoring, though, as with other assessment approaches, the extent of the match needed beyond subject matter is not clear.

Semi-structured Interviews

A limited number of combinations of teaching techniques, topics, and types of students can be specified in the tasks to be performed and discussed in the semi-structured interviews. In some of the tasks piloted, teachers were asked about how they might vary particular instruction for classes of "highly" and "less" capable students. Almost always, teachers interpreted this direction in light of their own teaching context, so the "highly" and "less" capable students envisioned varied considerably across teachers. For example, some inner-city teachers estimated that their "highly" capable students performed at approximately the 60th percentile in terms of national norms. Across all assessment approaches, teachers tended to respond in terms of their own students, despite explicit instructions to focus on a specific group of students which were not necessarily like their own.

Because of the limited ability to sample across teaching contexts, the ability of semi-structured interviews to measure a teacher's skills across contexts is limited. In addition, the interview responses are largely the teacher's description of what they might do, not what they actually do. While the interview might indicate a teacher's potential for applying skills, the relationship of a description of what might be done to actual application of skills in a classroom is not known. It is likely that other factors, such as the degree of classroom management skills or time pressures affect a teacher's ability to fully apply their knowledge in their classroom.

The tasks in the interviews vary from asking a teacher to compare and critique several instructional methods or techniques to asking a teacher to design a lesson for a specific purpose. In the latter case, the type of students and context can either be highly specified or not. If not, the task of designing a lesson allows a teacher to use the instructional techniques with which s/he is most familiar, providing that the task is appropriate for the types of students with which the teacher has had experience. The semi-structured interview provides ample opportunity for teachers to explain the rationale underlying their choices and to explain contextual factors which affected those choices. This facilitates the separate assessment of knowledge of content, of instructional strategies in the subject area, and of students. This separate measurement also facilitates diagnosis of specific areas of weakness, and may provide some indication as to whether or not a teacher is likely to succeed in a specific context.

However, in order to assess these separate domains, the skills within them need to be defined or indicated in such a way that the definitions apply equally well across various teaching contexts and teaching approaches. Interviewers and scorers would need to be trained using examples from teachers from different teaching contexts. It is possible that a match between the grade level, subject matter, and contextual experience of the scorer and the teacher being assessed might improve the accuracy of scoring.

Structured Simulation Tasks

As with semi-structured interviews, a limited number of combinations of types of students, teaching techniques, and topics can be presented in a set of tasks. This restricts the ability to measure the skills in teaching in diverse settings, although the sample is more broad than the two classroom-based forms of assessment (i.e., classroom observations and portfolios). As discussed earlier, however, the greater the simultaneous specification of students, topics and teaching techniques in the assessment materials, the more likely that a teacher has limited or no experience with one or more of these areas from which to draw in formulating a response.

The scoring methodology requires that the main characteristics of a correct response be specified prior to scoring of the structured simulation tasks. Therefore, this assessment approach, while able to represent the extent of a teacher's knowledge across a limited number of specific teaching situations, is limited in the ability to accommodate different conceptions of teaching. Teachers' skills in their own contexts may be underrepresented as the combination of topics, context, and instructional approaches with which the teacher has had experience may not be reflected on any of the tasks. Teachers in specialized contexts, such as continuation high schools and newcomer classes where students speak little English, criticized some of the instructional techniques and classroom assumptions represented in the tasks as not being relevant to their own experience.

In addition, some problems were experienced in the scoring of the pilot of the Structured Simulation Tasks for Secondary English Teachers in terms of different conceptions of teaching. These differences tended to be resolved by including specific characteristics of each teaching approach in the list of items in the lisc of response characteristics which received credit in the process of scoring. Therefore, a teacher giving an internally contradictory response could potentially earn many points. Secondary English is a particularly difficult subject area because there is not the degree of consensus about appropriate teaching approaches at the secondary level that exists in subject areas such as math and science. More thought needs to be given to how to accommodate different conceptions of teaching in the design and scoring of the tasks, since no resolution has been obtained as to the superiority of any approach in a given context.

Performance-based Assessment Center Exercises

Performance-based assessment center exercises, like other forms of simulation assessment, can only represent a limited number of combinations of students, teaching techniques, and topics in a set of exercises. This limits the ability to measure a teacher's skills in teaching in diverse settings. The ability to accommodate different conceptions of teaching varies according to the nature of the task, principally with the amount of discretion the teacher has in specifying the teaching approach and context represented in the response. For the reasons mentioned in the discussions of the other assessment approaches, the training and perhaps experience of the scorers continues to be important.

Videotaped Teaching Episodes

Again, a limited number of contexts can be portrayed in videotaped teaching episodes. Characteristics of a student or group of students in each context which affect instruction, if not obvious from the videotape, can be explicitly communicated through supplementary materials. Teachers can then be asked about the instruction of specific students or groups of students.

Videotaped teaching episodes can be extremely useful in testing a teacher's ability to recognize applied principles of effective instruction with specific groups of students, if the relevant student characteristics are either evident from the videotape or communicated to the teacher. In order to do this, however, careful thought needs to be given to the information given to the teachers being assessed and to the scoring framework employed. Some principles of effective instruction of an ethnically homogeneous classroom may be unknown to teachers, especially beginning teachers. For instance, methods of direct instruction resting on the teacher as authority prove to work well with black students (Hollins, 1982), but are counter to present trends in instruction which emphasize the teacher as facilitator.

The ability of the videotaped teaching episodes approach to accommodate different conceptions of teaching is unclear, as a method of conceptualizing and scoring the teacher responses to episodes which fully capitalizes on the video presentation has not been developed.

Multiple-choice Examinations

Multiple-choice examinations have the ability to sample situations broadly, and can portray multiple combinations of type of students, teaching techniques and topics within a subject area, either in the item stem or in the alternatives to be selected. In terms of knowledge of teaching in different settings, multiple-choice examinations can widely test knowledge of generally accepted principles of effective instruction of a variety of types of students. However, this assessment approach is extremely limited in its ability to measure the application of those principles. Since multiple-choice items must have a clear right or wrong answer, it is difficult to accommodate different conceptions of teaching in a single item. Knowledge of a variety of teaching approaches can be measured, through a corresponding variety of items; However, measuring the depth of knowledge and the ability to apply skills reflected in an individual teacher's conception of teaching in their own context through multiple-choice items is virtually impossible.

In general, the ability of any assessment approach to measure a beginning teacher's ability to teach diverse students is limited. This is due to sampling problems, the limited amount of experience possessed by beginning teachers, and to an absence of a knowledge base which integrates knowledge of the effective instruction of specific types of students into general principles which apply across groups.

Fairness across Groups of Teachers

Fairness across groups of teachers was difficult to ascertain because of the early stage of development of most of the assessment instruments piloted. With the exception of the CCI, instruments had not been developed to the point where reliable results were assured. The pilot tests focussed on identifying the strengths and weaknesses of the approach represented to inform any decisions for allocation of resources for further development. For this reason, only small numbers of teachers participated in the pilot tests, and the range of teachers represented was not great. An effort was made, however, to include in each assessment the following types of teachers: Males and females, minority teachers, teachers at all the different grade levels, inner city teachers, and rural teachers. Rural teachers were the most difficult group to include in the pilot tests because of their geographic dispersion and the costs -- both to the teachers and to the project -- of bringing teachers to more centralized locations. Despite the inclusion of different types of teachers, because of the small numbers of teachers participating in the pilot tests, group differences needed to be quite large to be statistically significant. Therefore, the potential effects on groups of teachers were indicated by qualitative responses on the evaluation forms from the various groups of teachers and patterns of differences in group means across tasks.

During the first year of pilot testing, there was very little time to recruit a wide variety of teacher participants. Therefore, no quantitative analysis of first-year results was done. Because of the greater lead time in the second and third years, a greater variety of teachers was obtained, though numbers were small. Quantitative results are available for the instruments pilot tested in the second and third year. Both qualitative and quantitative data are discussed for each instrument below.

For the most part, the vast majority of the participating teachers believed that each assessment instrument was appropriate across all groups of teachers. Generally, a few teachers evaluating each instrument expressed concern that teachers whose native language was not English might have difficulty in comprehending the instructions and explaining their ideas in English, especially in a timed test. Teachers in specialized contexts often expressed their frustration that the instructional techniques or tasks they were asked to perform for the assessment bore little relation to what they did in their classroom. These teachers taught in varying contexts. Some of these teachers taught newcomer classes, where new immigrants who speak little or no English are placed upon arrival in the district. Others taught in continuation high schools, where regular attendance of students is not common, and instruction must be designed to accommodate students who have missed previous instruction and who might not return for subsequent lessons. Teachers from districts relatively poor in resources reported a lack of familiarity with some instructional resources (e.g., equipment, films) specified in some tasks.

Because of the small number of teachers, quantitative results are only suggestive of possible group differences in performance on a fully developed assessment instrument. Generally, more pilot testing and analysis need to be done to obtain enough information to identify possible causes of group differences. However, if possible explanations for group differences were identified in the pilot testing, then they are discussed below.

Two types of data were used to identify possible group differences: (1) qualitative data from teacher evaluation forms, teacher comments during the assessment and staff observations of administration, and (2) quantitative data representing teacher performance on the assessments. Teacher comments and staff observations often identified difficulties that specific types of teachers were experiencing with the assessment, suggesting potential group differences in performance. Quantitative data was more problematic to interpret. In general, differences between mean scores of different teacher groups tended to be quite small and were not statistically significant. Therefore, differences in the same direction across multiple tasks, however small, were taken as a sign of possible group differences in assessment performance once the assessment approach is fully developed.

The comparison of assessment approaches is summarized in Table 5, with one column each representing conclusions drawn from quantitative and qualitative data, respectively. For more details, the relevant chapter analyzing the assessment instrument(s) in previous reports should be consulted. **As most instruments were in a stage of development that did not necessarily produce reliable results, and the number of teachers in each pilot test was almost always small, the following are working hypotheses only which should be more fully examined in field tests of any instrument which is fully developed.**

High-inference Classroom Observations

Passing rates of teacher groups were compared for both of the piloted instruments representing this assessment approach. For the CCI, which focussed mainly on general pedagogy, teachers at different grade levels were compared. High school teachers tended to perform less well than the middle/junior high school and elementary school teachers, especially in the areas of lesson content (e.g., accuracy and appropriateness of instructional content), lesson development, questioning techniques, and monitoring and adjusting instruction.

The Science Laboratory Assessment focussed on subject-specific pedagogy, general pedagogical skills, and knowledge of students. Analysis of group differences was complicated by ceiling effects; all but two of the twenty-nine teachers observed were judged as passing the assessment, and no teacher failed more than two of the seven domains. (Several teachers received no rating in one domain or more due to lack of sufficient information in those domains.) No clear patterns were observed in terms of teachers who received one or more failing ratings; groups of teachers compared included white/nonwhite, male/female, elementary/middle or junior high school/high school teachers, and teachers of varying age levels. As passing levels were not clearly defined, and there was some evidence of observers using different standards, therefore, the lack of group differences must be viewed with some caution.

Teachers generally believed the two assessment instruments to be fair across groups of teachers. Some features of the instruments piloted were designed to increase the fairness to some groups of teachers. For instance, the teaching competencies for the CCI were specifically defined in terms which were abstract enough to apply across teaching

TABLE 5

FAIRNESS ACROSS GROUPS OF TEACHERS

Assessment Approach	Groups for whom Assessment Results Suggest Potentially Large Differences	Other Concerns about Groups of Teachers
High-inference Classroom Observations	No trends observed in terms of age, gender, ethnicity, and grade level.	Instruction by vocational education teachers may not have the same characteristics as instruction by teachers in academic subjects.
Portfolios	Junior high school teachers did less well than high school teachers.	Teachers of LEP classes may do more oral activities which may result in difficulties in documentation.
Semi-structured Interviews	Insufficient data.	Teachers who have not previously taught the topics assessed may be at a disadvantage. Less verbal teachers may also be at a disadvantage.

TABLE 5 (cont'd)

FAIRNESS ACROSS GROUPS OF TEACHERS

Assessment Approach	Groups for whom Assessment Results Suggest Potentially Large Differences	Other Concerns about Groups of Teachers
Structured Simulation Tasks	Inner city teachers as a group had lower scores across tasks on both instruments than suburban and other urban teachers. White teachers as a group did slightly poorer across tasks than non-white teachers on one instrument, and did better on the other. There were no group differences across tasks for males/females and regularly prepared teachers vs. intern teachers. On one instrument, K-3 teachers as a group did better than 4-6 teachers.	Junior high school and middle school teachers were unfamiliar with some of the topics, which were more characteristic of high school instruction. Teachers of LEP students, teachers in continuation high schools, and teachers in poorer districts were unfamiliar with some of the materials, resources, and instructional approaches portrayed in the assessment. Some concern was expressed that the scoring system accommodate teachers from and of cultural groups with norms of strong adult leadership, as opposed to a more facilitating role.
Performance-based Assessment Center Exercises	Females as a group outscored males on all subparts. Suburban teachers as a group outscored urban/inner city teachers on all subparts. High school teachers as a group outscored middle school and junior high school teachers on 9 of 10 subparts. White teachers as a group outscored nonwhite teachers on 6 of 10 subparts.	A concern was expressed that the scoring system was biased against inner city teachers who deemphasized grammar and mechanics. Research on status differences in group participation was not considered in the design and scoring of the task centering on group discussion.

TABLE 5 (cont'd)

FAIRNESS ACROSS GROUPS OF TEACHERS

Assessment Approach	Groups for whom Assessment Results Suggest Potentially Large Differences	Other Concerns about Groups of Teachers
Videotaped Teaching Episodes	Scoring system was judged to be inappropriate in the prototype, making data unusable.	The teachers and students in the videotapes did not adequately represent the diversity in CA public schools. Cultural groups that have norms of strong adult leadership were not represented. The tasks separated instruction and management, which may be difficult for some teachers, e.g., Native Americans. Inner city teachers and black teachers may emphasize grammar and mechanics more than allowed in the scoring system.
Multiple-choice Examinations	Minority teachers performed less well than white teachers, except Asian teachers did equally well in science. Males outscored females on the science and social studies subtests, while the reverse was true for the pedagogy suhtest. Business, science, and social studies majors tended to outscore education, liberal arts, and English majors on all subtests.	Teachers who read relatively slowly seemed to tire near the end.

techniques, classroom contexts, and grade levels. Furthermore, some CCI teaching competencies are judged by patterns of student behavior (e.g., student engagement, student behavior suggesting the existence of classroom routines and norms for behavior.) The Science Laboratory Assessment also carefully matched observers to teachers in terms of experience at that grade level and/or area of science (e.g., physical science, life science.)

The fairness of this assessment approach, however, relies on the validity of the interpretations of teacher and student behavior made by the observer. Some effective behaviors, especially of teachers instructing an ethnically homogeneous group of students of the same ethnicity as the teacher, may be difficult for an observer to interpret. For example, sarcasm has been suggested as an appropriate and effective method of motivation and discipline for an African-American teacher teaching African-American students (Foster, 1989). Native American teachers instructing Native American students have been observed using facial expressions to effectively manage the classroom (Nelson-Barber and Meier, 1990.) As mentioned in the previous section, some reservations about the relevance of the competencies rated to effective instruction in vocational education were expressed by one of the observers.

Portfolios

The portfolio assessment focussed on a teacher's skills in teaching English at the secondary level, especially subject-specific skills (e.g., planning, subject-specific pedagogy) and the ability to reflect on one's own teaching. Each portfolio was rated by two scorers, who did not always agree in their ratings. All discrepant ratings were in adjacent rating categories, but some, ranging in percentage from 6 % to 31 % fell in the two rating points just above and just below passing. Therefore, scoring results are discussed in terms of "clearly passing" (i.e., both scorers gave a passing rating) and "clearly failing" (i.e., both scorers gave a failing rating.) Junior high/middle school teachers did less well than high school teachers. Of the six junior high/middle school teachers, four teachers did extremely poorly, with three clearly passing none of the six categories and one clearly passing only one category. In addition, all three of the teachers receiving their teacher preparation from out-of-state institutions were among the teachers receiving the lowest ratings, with one of these teachers commenting that their out-of-state experience was "almost groundless here."

All but one of the sixteen teachers believed the assessment to be fair across various teacher groups. Concern was expressed, however, that teachers of limited English proficient students might have more oral activities and hence would have more difficulty in documenting their teaching activities than teachers with few limited English proficient students.

Semi-structured Interviews

Little data for groups of teachers is available for the semi-structured interviews piloted. Data is available for the Semi-Structured Interview in Elementary Mathematics, but the scoring system was deemed inadequate. The scoring methodology for the Semi-Structured Interview in Secondary Mathematics (SSI-SM) and the Semi-Structured Interview in Secondary Social Studies (SSI-SSS) piloted a different, focussed, holistic scoring methodology which was in the process of being developed. Only two tasks administered for the SSI-SM were scored for a subset of the teachers. These tasks revealed inadequacies in the questions which were redesigned for a later version of the SSI-SM piloted with Connecticut teachers. The SSI-SSS scoring system, which was hoped to parallel that of the SSI-SM, had not yet been designed. Therefore, no quantitative data on teacher groups is available.

Although teachers were not aware of the scoring methodology when they commented on the assessment instruments, all of the SSI-SM teachers, 14 of the 16 SSI-SSS teachers, and 26 of the 41 SSI-EM teachers believed that the assessment in which they participated was fair across groups cf teachers. The discrepant findings may result from the peculiar nature of the SSI-EM, which was originally designed for expert teachers. Many of the participating teachers found it extremely difficult, both because it focussed on relatively sophisticated competencies and because they tended not to be well-grounded in mathematics.

The scoring of the semi-structured interview must be high-inference in nature because the teacher rationales can vary considerably in emphasis and content, and only a high-inference scoring system can accommodate these differences. As with other high-inference scoring systems, the fairness of the instrument rests upon the validity of the interpretations made by the scorers. If teachers talk about their own experience with their students, some degree of match between the teaching experience of the scorer and that of the teacher being evaluated is probably necessary, especially if diagnostic information is desired. Interviewer effects, (e.g., what happens when the interviewer and the teacher either are or are not of the same gender, ethnicity, etc.?) should be explored to determine any possible bias. Participating teachers in the pilot test strongly believed that teachers who had previously taught the material presented in the assessment were relatively advantaged compared to those who had not. Some also suggested that highly-verbal teachers might be favored over less verbal teachers because the assessment methodology put such a strong emphasis on explaining vs. doing. Some cultural norms (e.g., do not tell questioners what they already know; brief, succinct answers are preferable to more expansive ones) may cause the knowledge and skills of teachers in specific ethnic groups to be underrepresented in their interview responses. Similarly, people who are shy with adults or strangers may be at risk of underrepresentation of their knowledge and skills.

Structured Simulation Tasks

Two of the three pilot instruments representing this approach achieved sufficient reliability between scorers to warrant quantitative analysis of scores: The Secondary Life/General Science Teacher Assessment and the Assessment of Competence in Monitoring Student Achievement in the Classroom. The first focussed on secondary life and general science teachers, while the second focussed on elementary teachers In terms of comparison of mean scores on the assessment instruments, there were essentially no differences between the performances of males and females or, in the case of the secondary teachers, of teachers who were graduates of teacher preparation institutions and intern teachers. On both instruments, however, inner city teachers' scores were lower than non-inner city teachers. Performance differences between white and nonwhite teachers were mixed, with whites outperforming nonwhites on the elementary assessment instrument, and the reverse true for the secondary science instrument. For the elementary instrument, primary (K-3) teachers received higher scores than intermediate (4-6) teachers on three out of four test instruments.

The percentage of teachers who believed that the assessment instruments were fair across groups of teachers was high, ranging from 85 % to 95 % for each instrument. Concerns were expressed, however, about the topic, materials, and instructional strategies portrayed in the assessment materials. Teachers from poor districts, for example, commented that they couldn't afford the resources described in the tasks, and teachers of LEP students and continuation high school students said they used quite different teaching strategies than those portrayed in the assessment materials. In addition, Dr. Nelson-Barber expressed concern that the assessment materials might not reflect a variety of perspectives, and that some effective ways of teaching might not be recognized. For example, many effective black teachers emphasize strong adult leadership (Hollins, 1982; Delpit, 1988; Foster, 1989; Ladson-Billings, 1989), which contrasts with the more mainstream characterization of good teaching, which is seen as guiding and facilitating with the authority role deemphasized. Teachers designing lessons with more features of direct instruction might be penalized in the scoring criteria, if those criteria do not recognize direct instruction as an effective teaching method in some contexts.

Performance-based Assessment Center Exercises

Quantitative data on possible group differences in performance on performance-based assessment center exercises is dependent on just one instrument, so these data are even more tentative than those described for the other assessment approaches. Group differences in passing rates were noted on the single instrument piloted, the Secondary English Assessment. Females scored higher than males on all subparts and suburban teachers scored higher than urban/inner city teachers on all subparts. Though there were few minority teachers, white teachers outscored minority teachers on 6 of 10 subparts. High school teachers outscored middle/junior high teachers on 9 of 10 task subparts. Over half of the middle/junior high school teachers did not pass the task requiring them to respond to student writing. In particular, two-thirds of the teachers who did not pass the task measuring ability to discuss a literary work were middle/junior high school teachers.

Since this assessment was in the area of English, the previous caution about the effect of current trends in writing instruction which decreases the emphasis on grammar and mechanics in favor of an emphasis on the substance and structure of the writing should be heeded. This might put teachers of inner city students or teachers from educational backgrounds where grammar and mechanics are stressed at a disadvantage.

One exercise assigned an individual teacher score partially based on the teacher's participation in a group discussion of a literary work. It is incumbent on exercises where information is obtained through participation in groups to guard against the effects of status differences in group participation. Research has shown that members of lower status groups (particularly minorities, though sometimes women) participate in mixed-status groups less frequently than higher status members and, even when participating, their contributions are devalued by other members of the group (Berger, et al., 1974). While the discussion performance in this particular exercise was augmented by the teacher's notes on the work, scorers were not informed during preparation for scoring as to any rigorous procedure for recognizing and accounting for possible status differences in group participation.

Videotaped Teaching Episodes

The short-answer format employed with the single instrument pilot tested did not seem to be a valid way of measuring a teacher's knowledge for this assessment approach. Therefore, no quantitative data are reported for this approach.

Some potential problems with videotape as a stimulus medium were suggested which might result in differences between groups of teachers. Participating teachers expressed concern about the exclusive use of white, female teachers in the videotapes. Few students exhibited difficulties in understanding, and all students were quite fluent in English. Some participating teachers remarked on the differences between the students portrayed and their own students, who had serious learning difficulties and were often of limited English proficiency. Again, issues arose with respect to the limited representation of instructional strategies which ignored more authoritative forms of instruction and a deemphasis on grammar and mechanics.

Multiple-choice Examinations

The most common pattern for multiple-choice examinations is for members of nonwhite ethnic groups to score lower than whites (e.g., see results for California Basic Educational Skills Test (CBEST) in Watkins, 1985 and Majetic, 1990.) This pattern was also observed for the prototype instrument piloted, with the sole exception that Asians scored as high as the white group on the science subtest. Gender differences in performance on the prototype piloted were small, except for males scoring noticeably above females on the Science and Social Studies subtests and slightly below females on the Pedagogy subtest. The relative order of performance level across majors was roughly the same across all the five specific subtest areas, i.e., business, science, and social studies majors as groups scored above the mean on all subtests, and Education, liberal arts, and English majors scored relatively low on all subtests. The overall competence level made a relatively large contribution to performance across all of the subtests. For example, the score on the math subtest correlated well with the Math subscore on the SAT and with the score on the Science subtest, but correlated even better with scores on the Language Arts, Pedagogy, and Social Studies subtests.

The materials-based items employed in the instrument piloted greatly increased the amount of reading in the test, making it a test of reading ability as well as of knowledge of instruction. Although teachers were generally given as much time as they needed to complete the test, teachers who were slow readers seemed to tire toward the end.

COSTS

The costs of administering an assessment are dependent upon a number of factors, foremost among them being the complexity of the administration and scoring procedures for the assessment. This section presents an overview of the administrative complexity of each assessment approach, followed by an overview of the scoring complexity of each approach. Cost estimates for administration are also provided. As the costs of administering an assessment are also dependent upon that assessment being a completely developed product, this section concludes with a discussion of the degree of developmental work needed for the assessment instruments representing each assessment approach.

Administrative Complexity

Assessment approaches vary in administrative complexity, which affects costs. Summaries of descriptions of the administrative format and the complexity of administrative arrangements for each assessment approach pilot tested appear in Table 6. Each assessment approach will be discussed briefly to elaborate on the summaries provided in the table.

High-inference Classroom Observations

These require complex administrative arrangements. In addition to scheduling an observation time amenable to both the observer and the teacher observed, time must be scheduled for a pre- and/or post-conference. (Such conferences are necessary because of the high-inference nature of the assessment). Since a teacher's free time is limited, these conferences are often scheduled before or after school or during break times, necessitating the observer to spend more time at the school site than the time actually required for the observation. Consideration also must be given to scheduling observations so that there is a match, at least to some degree, between the teaching backgrounds of the observer and the observed. For example, it is probably best that an observer who has experience teaching secondary physical science be scheduled to observe a secondary physical science teacher. Similarly, an observer with experience teaching in inner city school contexts is probably best qualified to observe an inner city teacher. Furthermore, because observations are individually administered in a teacher's own classroom, arrangements must be made for the observer to travel to the teacher's school to conduct the assessment. For schools located in very rural settings, these travel arrangements can be quite extensive. Finally, because high-inference classroom observations entail a very labor intensive scoring system, and because the person who conducts the observation also scores it, observation schedules have to be

TABLE 6

FORMAT AND DEGREE OF COMPLEXITY OF ASSESSMENT ADMINISTRATION

Assessment Approach	Administrative Format	Degree of Complexity of Arrangements	Amount of Training Required for Assessors	Time Required Per Teacher
High-inference Classroom Observations	Individual	High, due to scheduling requirements	Extensive, as assessors also score	Averages about 1/2 day, counting travel time
Portfolios	Individual; all contact can be by phone and mail	Moderate, due to need to provide contact person to answer questions and to check portfolio for completeness		
Semi-structured Interview	Individual	High, due to scheduling and technology involved	Moderate, to standardize administration	1/2 day for one interview
Structured Simulation Tasks	Large group	Low	Minimal, consisting of standardized procedures to handle common problems.	Prototypes varied from 2 to 5 hours

TABLE 6 (cont'd)

FORMAT AND DEGREE OF COMPLEXITY OF ASSESSMENT ADMINISTRATION

Assessment Approach	Administrative Format	Degree of Complexity of Arrangements	Amount of Training Required for Assessors	Time Required Per Teacher
Performance-based Assessment Center Exercises	Depends on individual exercise; could vary from individual to large group	High, due to scheduling	Varies by individual exercise, depending on how much judgement is required	Varies by individual exercises; those piloted averaged one hour per exercise
Videotaped Teaching Episodes	Medium to large group	Moderate, due to technology involved	Minimal, consisting of standardized procedures to handle common problems	1/2 day
Multiple-choice Examinations	Large group	Low	Minimal, consisting of standardized procedures to handle common problems	1-4 hours

designed in such a way as to minimize mental or physical exhaustion on the part of the observer (i.e., generally only one observation should be scheduled for an observer per day).

If more than one observation is conducted over a period of time (e.g., a school year), then the administration arrangements necessary for one observation will be multiplied by the number of observations made. The complexity of arrangements may also be affected if all of the observations are not conducted by the same person (e.g., one observer may require more extensive travel arrangements than another).

Since observers also score the assessment, the training for administration is very complex, and is discussed in the section on scoring.

Portfolios

This form of assessment is moderately complex in terms of administrative arrangements. Each teacher must be provided detailed directions as to what is to be included in the portfolio, and each teacher should have access to a contact person familiar with the portfolio assessment with whom they can talk about the portfolio, problems they are experiencing, etc. Arrangements also have to be made for the submission and storage of portfolios, and for the provision of someone to check the portfolios as they are received to see if all components have been included. (This latter point is especially noteworthy because our pilot test experience revealed that even if teachers are provided binders with labelled dividers corresponding to the portfolio components and even if extensive directions are provided about each of the required components, teachers are still apt to submit portfolios with major pieces missing.) Consideration also needs to be given to "administering" the portfolio assessment over a long enough period of time so that teachers are not penalized for circumstances which affect their portfolio but are not under their control (e.g., a teacher has half of his/her students transferred out of the class midway through a unit or a semester).

Semi-structured Interviews

Like the classroom observations, the semi-structured interviews require fairly complex administrative arrangements. Although the interviews are individually administered, each task that is part of the interview is usually administered by a different assessor. Furthermore, because of the nature of an interview -- i.e., the interviewer must be able to hear the interviewee and vice versa -- each task and accompanying interview must be administered in a setting with a minimum of interfering noise. To help insure that the teachers being interviewed are not overheard by other teachers, a setting that provides for privacy should also be selected. Thus, ideally, each task and accompanying interview should be scheduled for a separate room. In addition, each task and interview should be scheduled so as to facilitate smooth transition between tasks. As it is difficult to equalize the time required for all tasks, and some teachers will finish a task more quickly than other teachers, consideration needs to be given to making arrangements to accommodate those teachers who finish a task early. Obviously, the above arrangements severely limit the number of teachers who can be assessed at a single administration.

Another factor contributing to the administrative complexity of this assessment approach is the technology currently used to document interviews (either videotapes or tape recorders). Precautions against the potential malfunction of equipment must be taken (even then, problems are likely to be experienced with some recordings with respect to quality of sound). The use of videotapes especially requires special arrangements as someone must set up and operate the equipment. If interviewers are responsible for the videotaping, some training usually needs to be administered in this area.

Training is also an important part of administration if provision is made for probing (i.e., follow-up questions to improve the clarity of responses and interruptions to redirect teachers who stray from the question asked). Assessors need to be trained to recognize when follow-up questions and redirection are required. Even if assessors are not allowed to probe but are required instead to stick to previously scripted questions, they must be trained to administer the interviews in a standardized manner and in a way that maximizes a teacher's chances of finishing the tasks and accompanying interviews within the set time limits.

Structured Simulation Tasks

This form of assessment has a low degree of administrative complexity because it can be administered in large group settings by one person who needs almost no training (training usually consists of standardized procedures to handle common problems), and it involves no technology. Depending on the type and number of stimulus materials used, each individual may require a fair amount of workspace, thus reducing the number of teachers that can be assessed at a single administration in a given amount of space.

Performance-based Assessment Center Exercises

This form of assessment is difficult to generalize about, as exercises can vary widely in terms of administrative complexity. Some exercises may be administered by one person, others require at least two assessors. Since teachers rotate from exercise to exercise, however, administrative logistics are complex in terms of scheduling, especially when formats and the time required vary between exercises. The administrative format of the exercises also limits the number of teachers who can be assessed with a single administration of the assessment.

The amount of training required for assessors also varies from exercise to exercise; the more the administration is not standardized but relies on professional judgements, the more extensive the training required.

Videotaped Teaching Episodes

These can also be administered to relatively large groups; however, the size of the group is limited by the number of video monitors available and the room arrangements necessary so that each teacher has a clear view of the video monitor and the sound is audible. Like semi-structured interviews, some sort of precautions must be taken to guard against the possibility of equipment failure, the occurrence of which would likely result in the cancellation of the entire assessment administration.

Depending on the technology arrangement, one person should be able to administer this assessment with a minimum amount of training in standardized procedures, using a VCR, and, possibly, in handling potential equipment problems. (Typically, equipment problems and installation of the video monitors are handled by a video technician who initially tests the system and then remains on call in case of problems.)

Multiple-choice Examinations

Like the structured simulation tasks, this form of assessment has a low degree of administrative complexity because it can be administered in large group settings by one person who needs almost no training (training usually consists of standardized procedures to handle common problems), and no technology is involved.

Complexity of Scoring Approaches

The way in which an assessment is scored directly affects the cost of administering that assessment. The assessment instruments pilot tested utilized a number of variations of two different scoring approaches:

> **Analytic** -- In analytic scoring, a teacher's responses are compared to predetermined criteria (i.e., answers). For every match between a response and a criterion, at least one point is awarded toward a total score, and in some cases, points are deducted for wrong answers.
>
> **Holistic** -- In holistic scoring, a teacher's responses are judged as a whole. Scorers use professional expertise to evaluate the quality of the response as a whole. Holistic ratings are often determined by comparing one teacher's responses with a set of criterion responses (which are sometimes the responses from other teachers). In addition, there are usually anchor responses established at least at the high and low end of a selected rating scale (e.g., 1 to 5), and often for every point of the scale.

The scoring systems for each assessment approach are summarized in Table 7 and discussed below.

High-inference Classroom Observations

Unlike many assessments, high-inference classroom observations are both administered and scored by the same person. The assessor is trained to first document the observation (usually some form of scripting) and then to reliably interpret and rate the observed teacher behaviors. Since the assessment methodology of high-inference classroom observations relies on professional interpretations of behavior of new teachers and their students, and as experienced teachers often approach assessment of new teachers with a personal and/or professional agenda for change, the training component of this assessment approach is extremely important so as to eliminate as much as possible all personal bias in the documentation and the scoring. Training is even more important when there are large differences between the teaching context and topic being observed and the observer's experience.

TABLE 7

FORMAT AND DEGREE OF COMPLEXITY OF SCORING

Assessment Approach	Scoring Format	Degree of Complexity of Scoring	Amount of Training Required for Scorers	Number of Scorers and Time Required Per Teacher for Scoring
High-inference Classroom Observations	Scripting or notetaking during observation, followed by categorization of notes, and evaluation of evidence.	High, due to breadth and depth of training required and time required to score.	Extensive, as assessors also score. The more distinct the indicators that are rated, the longer the training required.	One scorer per teacher; 4-6 hours
Portfolios	Holistic	Extensive, due to breadth and depth of training required.	Extensive, as scorers must reliably rate competencies across varying topics and grade levels.	Two per teacher; 1-2 hours per portfolio
Semi-structured Interview	Holistic comparison of performances to "marker tapes."	High, due to time required to calibrate scorers and train to record evidence.	Extensive, to orient scorers to anchor tapes defining performance levels.	Two scorers per teacher; roughly one day for an interview consisting of 4-5 tasks

105

106

TABLE 7 (cont'd)

FORMAT AND DEGREE OF COMPLEXITY OF SCORING

Assessment Approach	Scoring Format	Degree of Complexity of Scoring	Amount of Training Required for Scorers	Number of Scorers and Time Required Per Teacher for Scoring
Structured Simulation Tasks	Comparison of answers to previously determined list of ɐ.iswers; scorers use their professional judgement to award or deduct points	Moderate, due to need to calibrate scorers	Roughly 1/2 to one day per task, depending on the number of subparts and complexity of judgements to be made.	Two per teacher; roughly 10 minutes per task
Performance-based Assessment Center Exercises	Depends on individual exercises; those piloted were scored holistically	Complexity of actual scoring depends on the individual exercise; since it is iikely that each exercise is scored separately, scoring logistics are complex	Depends on individual exercises	Depends on individual exercises
Videotaped Teaching Episodes	Comparison of answers to previously determined list of answers; scorers use their professional judgement to award or deduct points	Unknown, due to lack of good model for scoring system	Unknown, due to lack of good model for scoring system	Unknown, due to lack of good model for scoring system

TABLE 7 (cont'd)

FORMAT AND DEGREE OF COMPLEXITY OF SCORING

Assessment Approach	Scoring Format	Degree of Complexity of Scoring	Amount of Training Required for Scorers	Number of Scorers and Time Required Per Teacher for Scoring
Multiple-choice Examinations	Machine scorable	Low, if technology is available	N/A; a machine is programmed with correct answers.	No human scorers required; machines can score multiple teachers per minute

To help eliminate the potential of bias, training for this assessment approach is extensive, usually covering a minimum of five days. Observers must be trained so that they have a common understanding of the competencies to be assessed and the rating standards used, as well as be trained in how to objectively document an observation, how to objectively interpret the data, and how to award a rating. Training provides observers with multiple examples of teaching behaviors in different contexts and gives them an opportunity to understand how their own ideas about teaching either do or do not conform to state expectations. The number of examples needed to effectively communicate the competencies assessed and the rating standards used is directly related to the number of distinct aspects of teaching covered. The greater the number of competencies and standards, the longer the training required. While this reliance on professional judgment requires lengthy training to achieve validity and reliability, it also provides the flexibility required to evaluate quite different styles of teaching in quite different contexts.

The actual process of interpreting and scoring the documentation from a high-inference observation assessment is usually a lengthy one, entailing from between three to six hours of an observer's time. Thus, the administration coupled with the scoring of one high-inference observation assessment usually "costs" one day of an assessor's time.

Portfolios

Portfolios are typically scored using a holistic approach. The amount of training needed for scorers would depend on the number and types of teaching skills to be assessed by the portfolio components; the greater the number and types of skills, the longer the training required. As with classroom observations, portfolios document a teacher's actual classroom and teaching context; thus, it is extremely important for scorers to be trained in the legitimacy of different teaching behaviors in different contexts.

As was the case with some of the other pilot tested assessments , each portfolio was scored by two people. As with the other approaches, the use of more than one person to score each assessment exercise would affect the estimated costs.

Semi-structured Interviews

Like high-inference classroom observations, semi-structured interviews require extensive scorer training to reliably interpret and rate teacher responses in a consistent manner. The most promising scoring approach developed to date is holistic in nature: One teacher's interview is identified as the standard for each point of the selected rating scale, and all other teacher interviews are compared to this set of anchor tapes (the interviews have been audiotaped or videotaped). The rating of the anchor tape is the rating given to those teacher interviews that most closely resemble the performance depicted by the anchor tape.

Training for this scoring approach is presently in the process of being refined. Since holistic rating methodology is highly developed with respect to writing assessment, it is likely, but not certain, that this scoring methodology can be developed to similar levels of reliability for these types of assessments. For the pilot tested assessments , scoring training lasted eight days per interview (approximately two days per task). Scorers were generally not the same people who administered the interviews.

Structured Simulation Tasks

Structured simulation tasks are usually designed with pre-determined correct responses, but not in a multiple-choice format. Instead, the teacher is asked to respond to the tasks in writing, the form of which can range from one paragraph to an outline of numerous pages. Thus, the assessment's pre-determined set of correct responses does not represent the universe of potentially correct responses, because it would be impossible to anticipate all of them. As scorers who have expertise in the subject matter and subject-specific pedagogical skills being assessed are usually recruited to score the teacher responses, some training is necessary to calibrate the scorers, i.e., orient them to recognizing the variety of forms which a correct response may take. The more judgement required (e.g., scorers may be asked to judge the depth of a response as well as its appropriateness), the more extensive the training required.

In some of the pilot tested structured simulation tasks, the participating scorers decided that instead of the analytic scoring method, a holistic scoring method was more suitable for some of the tasks. If this assessment approach includes both analytic and holistic scoring, then the scoring training would possibly be more extensive. Also, all tasks pilot tested were scored by two scorers each. Using more than one scorer for holistic scoring is typical, and so the costs of having more than one person score each exercise may need to be factored into the costs for this assessment approach.

Performance-based Assessment Center Exercises

The scoring of performance-based assessment center exercises varies with the individual exercise. While those piloted were scored holistically, it is possible to imagine other types of scoring (e.g., that employed for some of the structured simulation tasks). As each exercise measures distinct skills, it is virtually certain that training for each task will need to be conducted independently, assuring that scoring logistics will be complex. The amount of training for each individual exercise will vary with the degree of professional judgment demanded from the scorers; the higher the dependence on scorer judgment, the more lengthy the training required to calibrate scorers and the more time required to score each exercise.

As with the structured simulation tasks pilot tested, each of the performance-based assessment center exercises that were pilot tested was scored by two people. The use of more than one person to score each assessment exercise would affect the estimated costs for this assessment approach.

Videotaped Teaching Episodes

The prototype for the videotaped teaching episodes used a scoring system similar to those of the structured simulation tasks, where teachers provided short answer responses which were compared to responses previously determined to be correct. This methodology received much criticism from the scorers, and does not seem to fully take advantage of the strengths of the videotape methodology. Therefore, at this time, since a good example of a scoring system for videotaped teaching episodes has not been located for pilot testing, it is impossible to estimate the costs associated with scoring this type of assessment.

Multiple-choice Examinations

Epitomizing the analytic approach, the scoring process for multiple-choice examinations awards or deducts a point for a match between a teacher's response to an item and a previously determined correct response. These assessments are designed to be machine scorable (i.e., teachers are required to choose from among several choices and mark their answer accordingly), so there is no scoring training required. Programming the machine to recognize the correct answers is required, however, whereupon the score sheets can be scored extremely quickly.

Cost Estimates for Administration

Table 8 displays the costs estimates for administration and scoring of the assessment approach, based on the pilot testing experience. In calculating personnel costs, a standard rate of $20 per hour was assumed; estimates of personnel time needed were based on pilot

TABLE 8

COST ESTIMATES PROJECTED FROM PILOT TESTING EXPERIENCE FOR ASSESSMENT ADMINISTRATION AND SCORING

Assessment Approach	Per Teacher Cost Estimates
High-inference Classroom Observations	$134 - $157 per observation
Portfolios	$124 per portfolio
Semi-structured Interviews	$137 per half-day interview
Structured Simulation Tasks	$35 (two-hour) to $78 (half-day) per set of tasks
Performance-based Assessment Center Exercises	Varies with nature of exercises; estimate for those piloted was $133 per assessment.
Videotaped Teaching Episodes	Unknown, due to major changes anticipated in administration and scoring.
Multiple-choice Examinations	$32 - $40 per examination

testing experience. These costs do not include the cost of developing specific assessment instruments and managing the assessment system. The system design and the degree to which various assessment approaches might be merged with other systems would affect the management and related costs. These data do provide a rough basis of comparison between assessment approaches, when administered on a standardized basis.

Some of these assessment approaches, such as high-inference classroom observations and semi-structured interviews, might be used on a more informal basis by providers of support to new teachers to diagnose a new teacher's strengths and weaknesses. For instance, a support provider might use a semi-structured interview approach and ask a teacher to explain the rationale underlying a proposed lesson or unit plan.

Degree of Developmental Work Needed

Table 9 summarizes the stage of development of each assessment approach. based on the pilot testing experience and our limited familiarity with other instruments. Each assessment approach will be discussed separately.

High-inference Classroom Observations

High-inference classroom observations draw upon more than a decade of experience in using observations in the credentialing of beginning teachers, so little further development work is needed for this assessment approach methodology. In high-inference observations where assessors go beyond simply documenting the frequency of teaching behaviors to evaluate the appropriateness of the behaviors, it is important that the assessor understand the classroom context insofar as it affects instruction as well as a teacher's instructional goals for the particular lesson observed. The two piloted high-inference classroom observation instruments established procedures to facilitate such an understanding. First, the teacher completes a pre-observation information form describing lesson goals and activities and the students in the classroom. The assessor then discusses the information form with the teacher in a pre-observation interview to check the understanding of the classroom context and lesson goals. Procedures for both piloted instruments exhibited flexibility in accommodating any changes in the lesson needed to adjust for unanticipated events; any rationale for doing so was obtained through a post-observation interview.

Training of assessors consisted of communicating both the specificity and variety in application of competencies to be assessed. Methods used included guided discussions of the

TABLE 9

STAGE OF DEVELOPMENT BY ASSESSMENT APPROACH

Assessment Approach	Development Work Needed
High-inference Classroom Observations	Exploration of the extent of the generalizability of results for an individual, especially across topics.
Portfolio	Improved directions. Streamlined and better defined scoring criteria with performance markers for each rating point. More extensive training of scorers needed.
Semi-structured Interviews	Exploration of possible effects of interviewer differences and adaptation of training to minimize these differences. Process of summarizing evidence needs to be more standardized. To significantly reduce costs, scoring needs to occur simultaneously with administration.
Structured Simulation Tasks	More extensive pre-testing of materials, including directions and scoring protocols. Exploration of the use of analytic vs. holistic scoring methods. Design of tasks to accommodate conflicting views of effective instruction in a field.
Performance-based Assessment Center Exercises	Varies with individual exercise.
Videotaped Teaching Episodes	Identification of most appropriate uses of videotape technology as a stimulus for assessment of teachers. Identification and development of scoring methodologies other than short-answer responses.
Multiple-choice Examinations	Already well-developed, with well-established advantages and disadvantages.

competencies to be assessed, followed by multiple videotapes of teaching episodes to illustrate each competency and discussions focussing on the rating of a small number of competencies. Assessor proficiency was both achieved and checked through evaluating additional videotapes. Consensus with respect to skills in general pedagogy seemed easier to achieve than consensus on specific skills with respect to instruction in a particular subject, in this case, science. The definition of subject-specific skills was especially problematic across widely differing grade levels, e.g., elementary vs. secondary.

Although some data with respect to interrater reliability was collected for the more developed training system of the CCI, it was not made available for CNTP analysis. CCI assessors did pass a standardized proficiency test (consisting of rating a previously rated videotape) to establish their proficiency as assessors. In addition, reliability across observations of a single teacher is of concern, especially if the purpose of the assessment is to evaluate usual practice and not best practice.

Portfolios

High-quality portfolio assessments are characterized by (1) clear instructions for the selection and/or development of portfolio entries; (2) a scoring process which includes specific criteria by which a portfolio is to be evaluated which are appropriate for beginning teachers; and (3) extensive training for scorers which focusses on the ability to evaluate each criterion for a variety of contexts and topics. The piloted proto' pe was not always successful in eliciting the desired portfolio entries from participating beginning teachers. Despite an orientation handbook and the provision of binders with clearly labelled sections indicating portfolio entries, some teachers returned incomplete portfolios or portfolios with entries that did not satisfy the described requirements. Teachers reported difficulties in assembling student work with teacher responses, particularly in collecting a complete set of work of a student over time. Access to xerox machines was problematic for some teachers, and others found the expense of duplicating student work to be a burden. Methods of documenting oral activities also need to be developed. Some teachers reported frequent use of oral activities to both instruct and evaluate their students; therefore, they did not believe that the portfolio accurately documented their teaching. This was especially true of teachers of limited-English proficient students. These problems in administration need attention before any statewide implementation of a portfolio assessment.

Scoring criteria should clearly focus on teaching skills; some criteria used in the piloted prototype more closely measured the ability to follow directions. The appropriateness of any proposed criteria for beginning teachers should also be examined, as approximately one-quarter of the teachers participating in the pilot test did not pass any of the six categories used to score the piloted prototype.

Holistic scoring proved to be an effective way of evaluating the portfolios, allowing flexibility across a variety of topics and teaching contexts. The training methodology used in the pilot test seemed workable, but would be significantly improved by: (1) the development of a sample portfolio for use as a model in training; (2) the development of a scoring guide which includes performance markers for different ratings for each criterion to be evaluated; and (3) the lengthening of training to provide greater opportunities for discussion and practice using the scoring response criteria and form. Portfolio scorers need to be experienced in teaching the subject being assessed, familiar with focussed holistic scoring, and knowledgeable about a variety of teaching contexts.

Semi-structured Interviews

Important features of training for interviewers in the administration of semi-structured interviews includes reviewing the interview protocol, including the identification of any anticipated problems and possible solutions, tips in establishing rapport, and standardized ways of asking questions and constructing additional probes. Videotaped practice interviews coupled with specific feedback on the performance were cited as extremely helpful by the interviewers who participated in the training. However, even with this extensive preparation, observation of the interviews revealed variance between interviewers in tone (friendly vs. formal) and in the frequency with which they asked probing questions. Before semi-structured interviews are adopted as a standardized assessment approach, the effects of interviewer differences with respect to possible biases introduced need to be explored, and training needs to be modified to minimize these differences.

During the period when the semi-structured interviews were being analyzed, Connecticut assessment developers tried several methods of scoring semi-structured interviews. Analytic techniques were finally abandoned in favor of a focussed holistic approach. This approach uses actual teacher performances and accompanying documentation to define rating categories. This approach seems far superior to the analytic approaches used previously by Connecticut staff and by the Stanford Teacher Assessment Project. Analytic approaches had great difficulty in identifying categories or indicators that satisfactorily reflected differences in the complexity of teacher thinking and at the same time applied equally well across all responses. The training process, though lengthy, was able to produce moderate correlations between scorer pairs rating the same teacher. The process of recording information from the videotapes and arriving at judgements seems to be well developed. Less satisfactory is the process of producing a summary documentation of the judgment reached: A process which results in more standardization of summary evaluation forms, e.g., the use of some standardized language or some standardized categories, needs to be developed.

Teachers were rated at an indicator level, with five indicators per task, two scored tasks per topic, and two topics. Results from an early version of the holistic scoring system found moderate inter-rater correlations. Coefficient alphas, a measure of internal consistency, were moderate to high for aggregated scores across all indicators, across all indicators within a task, all indicators within a topic, and across indicators.

A semi-structured interview in a different topic, social studies, was piloted with initial work begun on developing a parallel scoring system. The set of indicators used for the math interview focussed on content knowledge (concepts, basic principles), content/curriculum (relationships between concepts), content pedagogy, general pedagogy, knowledge of student backgrounds and interests, and knowledge of student abilities. This system seemed transferable to the social studies interview, with one exception. The mathematics interview used two indicators of content knowledge: One focusing on definitions of concepts and correct use of terminology, and another focussing on the relationships between concepts. In considering the use of these two indicators for the social science interview, the assessment developers decided that a major focus of the content of social studies is on relationships between concepts and between concepts and themes. The resulting difficulty in collecting separate evidence for the two indicators of content knowledge led to a decision to combine them into a single indicator for use in the social studies interview.

Finally, to significantly reduce costs, reliable scoring needs to occur simultaneously with administration, a feature yet to be developed. This might be technically feasible after the semi-structured interview assessment approach is further explored.

Structured Simulation Tasks

Three sets of structured simulation tasks were piloted, with two being significantly advanced in terms of development, compared to the third. The two sets exhibiting the most development were based on a methodology currently used in the applied practice portion of some bar examinations used to license lawyers. This methodology has proved of sufficient technical quality to use in the legal profession, and needs further adaptation to teaching.

Tasks which required teachers to list general principles with little reference to the prompting problem, e.g., listing four characteristics an observation assessment should have, were judged to reflect more a teacher's ability to memorize than their ability to apply knowledge. Other tasks, e.g., construct three multiple-choice items based on a given passage, were judged to be more authentic measures of a teacher's knowledge, skills, and abilities. Three major developmental needs were identified in the pilot test of the prototypic structured simulation tasks: (1) improvement of the clarity of directions and scoring rubrics, some of which were abandoned or greatly changed during the course of scoring; (2) refinement of the scoring methodology used; and (3) development of a process for creating scoring criteria when there is controversy in the field over the effectiveness of specific instructional approaches.

In the set of structured simulation tasks for science teachers, instances of unclear directions were encountered which resulted in either scorers discarding a portion of a task or in penalizing a number of teachers who misinterpreted the directions. These tasks were not pretested prior to the pilot test administration. However, the English assessment materials were pretested with roughly ten beginning teachers, and problems were still identified with those materials. Some problems, e.g., lack of clarity as to what is expected in the way of a response, might be solved with more extensive orientation materials which provide examples of some tasks and teacher responses. Other problems center around requests for additional information which the teachers believed they needed to respond appropriately. For instance, in a lesson planning task, teachers were instructed to design a series of lessons to prepare students to write a compare/contrast essay on a specific literary theme. The instructions did not indicate whether students had previously written this type of essay. Teachers who assumed that they had not and centered their instruction around teaching students this type of essay were penalized, as instruction was supposed to center around the literary theme. The pilot testing experience suggests the need for careful and perhaps extensive pretesting of assessment materials.

In both the science and the English structured simulation tasks, scoring rubrics initially proposed were abandoned during the course of scoring and replaced with substantially reconceptualized ones. In many cases, this meant exchanging a more analytic method for a holistic one. Scorers were sometimes skeptical whether the identification of a lot of small, isolated features indicated greater knowledge and ability with respect to performing a task, and preferred a holistic scoring method. The use of holistic scoring, however, increases the preparation and time required for scoring, reducing one of the major advantages of this methodology. As types of tasks are developed, the optimal scoring methodology should be identified for parallel versions of the same tasks. This should eliminate the radical changes in scoring methodology experienced during the pilot testing.

When holistic scoring is used in assessments which are to be comparable over time, the focussed holistic approach where already rated examples are provided to guide scorers is preferred to a norm referenced scoring system where the scoring standards differ with each set of papers being scored. Again, there are models of holistic scoring which can be drawn from writing assessments. Methods of checking for scorer drift, such as mixing already scored responses into the set of responses to be scored, are also available to check on scorer proficiency.

Perhaps the thorniest problem to be worked out is what to do when there are differing views within the field about what constitutes correct instruction. In approaches such as high-inference classroom observations, portfolios, and semi-structured interviews, the evaluation can be tailored to the teaching philosophy used by the teacher, checking to see that weaknesses in the resulting instructional activities are acknowledged and compensated for. The structured simulation task approach requires that salient characteristics of the response be specified in advance, limiting scoring flexibility. During the scoring of the tasks for English teachers, different and conflicting philosophies of the teaching of writing emerged among both the scoring development team and the scorers. Combining salient features of all philosophies into a single analytic scoring system weakened the validity of the resulting scores, as teachers who inappropriately combined elements of more than one philosophy could potentially score higher than teachers who thoughtfully applied a single philosophy. One solution might be to avoid such tasks until further research compels a consensus. However, the disagreement is over the appropriateness of instructional activities in tasks which are central to subject-specific pedagogy; avoiding these tasks weakens the validity of the assessment. A thoughtfully designed holistic scoring method might resolve this problem, but more exploration needs to be done to be certain.

Performance-based Assessment Center Exercises

The technical quality of each exercise is independent, and is likely to vary with the degree to which the exercise is innovative, i.e., does not rely upon previous experience in assessing the focal knowledge, skill or ability.

Videotaped Teaching Episodes

The piloted prototype for videotaped teaching episodes exhibited a number of problems. Questions sometimes could be answered without watching the videotaped events, failing to capitalize on the stimulus format. Internal inconsistencies in teacher answers suggested that the short-answer scoring format has validity problems. The major need in development of videotaped teaching episodes is to identify aspects of teaching which are best communicated through the medium of videotape, and to identify the best way to solicit demonstrations of a beginning teacher's knowledge, skills, or abilities using the videotape as a cue. The work of Berliner and his colleagues (1989) suggests that beginning teachers have difficulty interpreting videotaped lessons, so the construction of a videotaped teaching episodes assessment for beginning teachers is challenging. If teachers are asked to develop a product in response to a videotape, then an analytic scoring methodology similar to that used for structured simulation tasks might be appropriate. If teachers are to analyze or critique the videotapes, then a focussed holistic scoring methodology which captures the depth of their thinking might be appropriate.

Multiple-choice Examinations

Multiple-choice examinations are the product of decades of developmental work. Extremely sophisticated methods of measuring the technical quality of these types of instruments are available. The more materials-based items which were hoped to increase the ability of the multiple-choice examination assessment approach to measure application of knowledge, skills, and abilities did not seem to do so to any great extent. In addition, the items seemed to decrease the advantages of the approach without significantly addressing the disadvantages (i.e., the length of time for administering each item was substantially increased without a corresponding increase in assessment of depth of knowledge.) Without a further technological breakthrough in the development of multiple-choice tests, this assessment approach seems to be fully developed, with well-established advantages and disadvantages.

ANALYSIS OF TECHNICAL QUALITY

Most piloted instruments had serious weaknesses in technical quality, because of their early stage of development. The architects of the CNTP had to choose between exploring a wide variety of innovative assessment approaches in the early stages of development, or selecting fewer approaches and trying to develop them to the quality necessary for implementation. Since little was known about the potential of many of the assessment approaches, as they were relatively new, the former course was chosen. The sole exception to the weakness in technical quality was the CCI, a high-inference classroom observation instrument. The CCI had undergone several years of development, and drew from a decade of experience of other states with a series of classroom observation instruments, each of which tried to improve on those before. The remainder of the pilcted instruments represented attempts to explore the potential of a relatively new assessment approach or, in the case of multiple-choice examinations, to extend the capability of a well-established assessment approach. With the exception of videotaped teaching episodes, methodologies for the design, administration, and scoring of the assessment approaches which hold the promise of meeting standards for fairness, reliability, and validity have been developed and are available. However, some methodologies need substantial developmental work to solve remaining problems.

In reflecting the complexity of teaching, nearly every scoring methodology requires a high level of inference from the scorers. Even the more analytic methodologies employed in scoring the structured simulation tasks require inference from the scorers when matching the teacher responses to the predetermined criteria. The pilot testing observations and analysis revealed some principles underlying the design of high-inference instruments or holistic scoring systems which seem to apply across assessment approaches. Although the principles seem obvious, they were not followed in the design of all of the assessment instruments piloted. To validly and reliably measure a teacher's knowledge, skills, and abilities, the following principles must be followed:

- The knowledge, skills, and abilities being measured must be clearly conceptualized and defined through terms or examples which are specific enough to define the knowledge, skill, or ability being evaluated, but abstract enough to apply across teaching contexts and instructional approaches.

- A number of examples from varying contexts and instructional approaches to illustrate the rating of the knowledge, skills, or abilities must be provided during the training of the scorers and, if applicable, administrators. These examples should clearly communicate the knowledge, skill, or ability being measured at the same time as illustrating the diversity in application.

- When using high-inference scoring, matches between participating teachers and scorers are important. Matches with respect to subject matter, grade level, and context seem especially critical. The extent of the match needed to evaluate general pedagogical skills at the basic levels of competency expected of beginning teachers is not clear. It seems critical that as long as the scorers have experience in the type of instruction (e.g., cooperative learning, lecture) evaluated if students are homogenous with respect to a cultural group having significantly different norms for interaction from mainstream society, then experience with that type of students also becomes important. Evaluation of subject-matter knowledge and subject-specific pedagogy, and perhaps knowledge of students, require the highest degree of match.

The ability of assessment instruments to adhere to these principles is dependent upon the state of knowledge about the area of knowledge, skill, or ability being evaluated. In general, the highest degree of professional consensus among educators as to the key knowledge, skills, and abilities is in the area of general pedagogy. Subject-specific pedagogy varies with the subject area. For example, there appears to be greater consensus with respect to what is appropriate instruction in the areas of mathematics and science than in English. Knowledge of students is an uneven area in terms of widely accepted principles. Much is known about the instruction of students varying in terms of academic ability; little is known about general principles for successful instruction of students from different cultural groups, and what is known seems to defy generalization, either within or across groups. It is to be expected that the instruments will continue to improve as research and practice in teaching becomes further codified. Because teaching is so context dependent and involves multiple and contradictory goals (e.g., fostering student independence and teaching students how to work cooperatively and collaboratively), however, it is doubtful that the profession can ever be reduced to a small set of principles which can be applied simultaneously in every situation. Teaching seems to more resemble a series of tradeoffs between efforts to achieve multiple goals which frequently require conflicting strategies (Lampert, 1985). Even assessment instruments which reflect this complex vision of good teaching may differ in the quality and quantity of information which they provide and the degree to which that information is appropriate for various purposes.

Suitability for Different Purposes.

Each approach piloted, with the exception of videotaped teaching episodes, seems to have the ability to be developed to yield reliable and valid results. However, these results will not necessarily be equally appropriate for different types of decisions to be made with respect to data from one or more teachers. Four major purposes for use of data from standardized assessments include: (1) licensure/certification; (2) feedback to teacher preparation programs; (3) hiring and retention; and (4) professional development. First, general requirements of assessments will be discussed for each assessment approach; then each assessment approach will be discussed in terms of its appropriateness for each potential use.

Licensure/certification

For licensure and certification decisions, assessment instruments must be very accurate around the cut score, i.e., the borderline between granting and denying the license or certification. As discussed earlier, one policy decision with respect to teaching competency is whether to concentrate on the demonstration of skills in the context in which a teacher is actually teaching or on the demonstration of skills across a variety of settings. Given that beginning teachers have experience in an extremely limited number of contexts, an assessment policy using the latter approach must use standards reflecting less depth of knowledge and less complex skills than standards for an assessment policy using the former approach. Every assessment methodology entails sampling topics, types of students, and instructional strategies. Thus a teacher's rating on a particular simulation instrument will reflect the knowledge of the topic, type of student, and instructional strategy represented on the assessment instrument. Assessment instruments that are not classroom-based may underrepresent the knowledge, skills and abilities of teachers as applied in the particular classroom in which they teach.

A small study comparing the evaluations of teachers made by the piloted assessment instruments with evaluations made during student teaching and the first year of teaching is currently in the process of being conducted. Preliminary results suggest that at least some assessment instruments may measure a teacher's potential with respect to teaching skills, and not necessarily how they perform in the classroom. Beginning teachers may need some assistance to realize their potential, i.e., to successfully apply their knowledge of teaching and students. Some teachers appeared to do better in the classroom than suggested by their assessment evaluation, e.g., in designing a lesson. The data are not sufficient to identify the source of the discrepant evaluations, suggesting further study to explore the relationship between assessment results and performance in an actual classroom.

The classroom-based assessment approaches (e.g., high-inference classroom observations and portfolios) measure a teacher's ability with respect to the students and teaching context. Such instruments cannot measure a teacher's knowledge, skills, and abilities with respect to other students and other contexts. However, a beginning teacher's knowledge is probably greatest with respect to the types of students, topics, and instructional techniques with which they have had experience. Therefore, it is likely that the depth of knowledge, skills, and abilities represented in standards for beginning teachers can be greater for classroom-based assessment approaches than for simulation approaches (e.g., semi-structured interviews, structured simulation tasks, performance-based assessment center exercises) and other approaches (e.g., videotaped teaching episodes and multiple-choice examinations.)

If policymakers decide that it is a compelling State interest to see that beginning teachers have some knowledge, skills, and abilities with respect to a variety of students and topics, then simulation approaches and other approaches are preferable assessment approaches. However, the simulated performance will, at best, only approximate actual performance, and the exact relationship between the two needs exploration.

Feedback to Teacher Preparation Programs

Teacher preparation programs could potentially use information from assessments of beginning teachers in meeting Standard 6: Program Development and Evaluation of the Standards for Program Quality and Effectiveness for each credential program. To provide feedback on the program, scores would need to go beyond a simple pass/fail measure and be criterion-referenced instead of norm-referenced. Patterns of scores would potentially indicate strengths and weaknesses of each credential program. The utility of the information would depend on the congruence between the curriculum of the credential program and the teaching competencies being assessed. The utility would be greatest for those competencies which might be expected to be most fully developed during teacher preparation, i.e., those which do not require extensive experience to develop, such as establishment of routines and questioning techniques. Competencies which require more ex' .nsive experience, e.g., evaluation of student learning which requires some experience with student error patterns, will also be affected by the context and degree of support available in the first teaching position. Therefore, this type of competency will be less reflective of the effectiveness of the credential program which prepared a teacher.

Hiring and Retention

Local education agencies are less interested in a teacher's ability to perform across a variety of contexts and more interested in how a teacher will perform in the particular district context in which there is a vacancy to be filled. At the present time, there is insufficient experience with most assessment approaches to understand how generalizable the results are to different contexts and topics. Some districts already use some of the as^essment approaches (e.g., classroom observation, semi-structured interviews, structured simulation tasks) in evaluating applicants; however, these local assessments are generally low in technical quality, lacking explicit criteria and standards (Izu, et al., 1992).

Professional Development

For use in professional development, an assessment approach needs to clearly and accurately indicate a teacher's strengths and weaknesses in such a way as to indicate specific areas where assistance might be helpful. To do so, each domain of teaching must be broadly sampled. Some assessment approaches provide a substantial amount of information, but only for a few topics at most. Some assessment approaches (e.g., classroom observations, portfolios, semi-structured interviews about portfolio entries or materials the teacher is in the process of developing) have greater potential to identify strengths and weaknesses of a particular teacher. If the diagnosis results in individualized assistance, then frequent interaction between the beginning teacher and the support provider can allow the use of assessments which are lower in technical quality than those used for statewide purposes such as credentialing.

Assessment approaches vary in their correspondence to the technical requirements described above for various uses. The remainder of the section discusses the appropriateness of each assessment approach for each potential use, summarizing the conclusions in Table 10.

High-inference Classroom Observations

High-inference classroom observations are developed to the point that they could be used for licensing to verify that a teacher has successfully demonstrated specific skills with at least one group of students, particularly if general pedagogical skills are the major area of focus. However, they have three potential disadvantages for use in licensure: (1) they are one of the more expensive assessment approaches in terms of statewide administration, particularly if skills in subject-specific pedagogy are included; (2) the results may not be generalizable across contexts or topics; and (3) a teacher's performance on high-inference classroom observations is likely to be affected by the difficulty of the teaching assignment and by the degree of support available to a teacher in the beginning years.

Results of high-inference classroom observation assessments could provide useful feedback to teacher preparation programs. As with any assessment approach, in order to interpret aggregate results of graduates, faculty would need to be familiar with the criteria and standards employed, perhaps even trained in the use of the instrument. Provision of individual results would assist in factoring out the differences in level of support, if this is known for the relev nt districts. Information beyond pass/fail statistics would also be

TABLE 10

SUITABILITY OF ASSESSMENT APPROACHES FOR DIFFERENT PURPOSES

	PURPOSE			
Assessment Approach	**Licensure/ Certification**	**Feedback to Teacher Preparation**	**Hiring and Retention by Districts**	**Professional Development**
High-inference Classroom Observations	Useful in assessing ability to apply skills, but expensive. Results may not generalize to other topics and teaching contexts.	Potentially useful, if results beyond pass/fail statistics are available.	Results of standardized administration potentially useful in hiring, if criteria correspond to district priorities. Probably too time-consuming for local use in hiring by most districts. Main method currently used the evaluations for retention.	High potential for local use in the areas of instruction, classroom management, and classroom climate. Prompt feedback is desirable.
Portfolios	Useful in assessing the ability to apply skills, especially in the areas of planning instruction and evaluating students, but expensive. Results may not generalize to other topics and teaching contexts.	Potentially useful if results are provided in great detail. Information on reflectivity might be useful.	Results of standardized administration potentially useful for hiring, if criteria correspond to district priorities.	Results of standardized administration could be useful, especially in subject-specific pedagogy. High potential for local use, if time is available.

TABLE 10 (cont'd)

SUITABILITY OF ASSESSMENT APPROACHES FOR DIFFERENT PURPOSES

	PURPOSE			
Assessment Approach	**Licensure/ Certification**	**Feedback to Teacher Preparation**	**Hiring and Retention by Districts**	**Professional Development**
Semi-structured Interviews	Useful in assessing the depth of subject-specific pedagogical knowledge and knowledge of students, but expensive. Results may not generalize to other topics and teaching contexts.	Potentially useful in identifying patterns of weakness in subject-matter knowledge, but interpretation of other patterns might be problematic unless teacher has had experience teaching focal topics.	Limited utility of results from standardized administrations for hiring, because application skills are not well assessed. Some potential for use in retention decisions to measure subject-specific pedagogical knowledge, if topic of interview corresponds to teaching assignment.	Results from standardized administration may not generalize to other topics and teaching contexts. High potential for local use, if time is available.
Structured Simulation Tasks	High potential for assessing general competence in teaching particular subjects. Relatively inexpensive.	Limited utility, as no information beyond passing rates is available.	Limited utility, as no information beyond passing rates is available.	No feedback is available to guide staff development activities.
Performance-based Assessment Center Exercises	Potential depends on whether skills judged as important lend themselves to simulation. Expensive.	High potential, if quality of simulation is high.	Use of results from standardized administrations depends on generalizability of results. Too expensive for local use.	Depends on importance of focal skills to districts. Some depth of information available.

TABLE 10 (cont'd)

SUITABILITY OF ASSESSMENT APPROACHES FOR DIFFERENT PURPOSES

	PURPOSE			
Assessment Approach	Licensure/ Certification	Feedback to Teacher Preparation	Hiring and Retention by Districts	Professional Development
Videotaped Teaching Episodes	Not suitable without future methodological improvements.	Not suitable without future methodological advances.	Not suitable without future methodological advances.	Not suitable without future methodological advances.
Multiple-choice Examinations	Useful in measuring breadth of subject-matter knowledge, though weak in measuring depth. Also weak in measuring pedagogical skills. Relatively inexpensive.	Little use in measuring pedagogical skills.	Little use in measuring pedagogical skills.	Little use in measuring pedagogical skills.

helpful in identifying common weaknesses in graduates which might be addressed in teacher preparation.

At present, the results of high-inference classroom observations are available at the time of hiring to districts from student teaching evaluations, but tend to receive limited use. Districts vary in the salience of specific criteria; if at least some of these criteria were addressed, more standardized high-inference classroom observations could make an increased contribution to hiring decisions. Again, however, district personnel would need to be familiar with the criteria and standards in use, and find them appropriate. High-inference classroom observations are already the major form of assessment for retention decisions in districts, though criteria and standards are not always well-defined. There is some overlap between the CCI competencies evaluated and the most common skills evaluated by districts according to a study sampling California teacher assessment practices, although the CCI competencies are more clearly specified than most of those in use by districts. An observation instrument could be designed that focusses on other areas, e.g., subject-specific pedagogy, that are not currently well-represented in district assessment practices. However, instruction, classroom management, and classroom climate are the areas best assessed by classroom observation instruments, and any additional areas are probably better assessed by another assessment approach.

High-inference classroom observations have high potential to provide useful information to a beginning teacher concerning instruction, classroom management, and classroom climate. Since the evaluation addresses a teacher's actual teaching of students familiar students, the amount of interpretation of results by the teacher is reduced, making it more likely that the evaluation results will be understood. To do this, feedback beyond pass/fail results must be provided as soon as possible after the observation is performed.

Portfolios

With more developmental work, portfolios can meet the accuracy at cut points required of assessments for statewide licensure. A portfolio is probably strongest when it documents a series of lessons, when consistencies and contradictions of various entries become evident. Clear criteria and standards which span topics and contexts, similar to those of high-quality high-inference classroom observations, are required. Training to evaluate portfolios is, therefore, a lengthy process, as evaluators must be able to apply standards evenly across a wide variety of situations. Therefore, portfolios are likely to be one of the more expensive assessment approaches if administered statewide.

Like high-inference classroom observations, portfolios have high authenticity, requiring teachers to actually demonstrate skills and abilities with at least one group of actual students. However, they share all the disadvantages of high-inference classroom observations, such as generalizability. Portfolios are superior to high-inference classroom observations in evaluating skills in subject-specific pedagogy, but weaker at evaluating general pedagogical skills. However, portfolios require a substantial investment of time from a teacher, and beginning teachers already have high demands on their time. A statewide portfolio assessment should be delayed until after the first year of teaching. This additional time would also allow the teacher to acquire additional experience, and would facilitate the development of subject-specific pedagogical skills by enabling comparisons between experiences with different topics and students.

Portfolios have the potential to identify common strengths and weaknesses in graduates of teacher preparation programs. However, the strength of a portfolio is in evaluating skills in planning, instructional design, and diagnosis and evaluation of student learning. These are unlikely to be substantially developed without sufficient experience to identify student error patterns, topics where students commonly experience difficulty, and the time requirements of different instructional techniques. These skills are likely to be developed only to a limited extent in student teaching. If a portfolio requires some reflective component such as an essay where the teacher describes and analyzes the implementation of a lesson or unit, then some data on reflective ability is available, which can be developed in teacher preparation. As with classroom observations, the degree of support and difficulty of the teaching assignment will likely affect the performance of a beginning teacher, and individual-level information beyond pass/fail statistics is likely to be most useful.

Some teacher preparation programs are beginning to assist their student teachers in the development of portfolios for use in job interviews. The utility of portfolios to districts is not clear, as their evaluation typically requires at least a half-hour per teacher. Portfolios might be most useful to districts in the final stages of the hiring process when the applicant pool has been reduced. Portfolios can document aspects of teaching for purpose of retention that observations cannot, e.g., skills in planning instruction and evaluating students. Some districts require beginning teachers to submit lesson plans, a common portfolio entry, which are then reviewed. Like classroom observations, standardized portfolios would be useful to districts only insofar as the criteria and standards used resemble district priorities.

Portfolios can contribute to professional development decisions, indicating areas of strength and weakness, especially in the area of subject-specific pedagogy, which in a sample of districts studied often received little attention beyond whether or not the district curriculum was being followed. Again, more information beyond pass/fail results could be useful in directing staff development decisions. If they have the time, support providers may find portfolio entries useful in displaying a beginning teacher's skills and in evaluating progress in problematic areas. Portfolios will not be useful in diagnosing problems in classroom management or, in the absence of a self-reflective entry, in interactions with students.

Semi-structured Interviews

Semi-structured interviews are not now suitably accurate at the cut oints for use in licensure, though present inter-rater reliabilities are sufficiently high to suggest that further developmental work is likely to achieve the desired accuracy. Semi-structured interviews are expensive to administer and score; simultaneous administration and scoring would lower the costs considerably, but the labor-intensiveness of administration will keep costs high compared to less labor-intensive assessment approaches. A teacher's explanation of the planning and instructional strategies used allows the semi-structured interview to measure the depth of a teacher's knowledge, and some degree of skills in application of that knowledge, but the degree of relationship between performance on a semi-structured interview and ability to effectively apply the knowledge and skills exhibited is unknown. Semi-structured interviews are particularly good at measuring the depth of a teacher's skills in subject-matter pedagogy, and show some potential for measuring knowledge of students.

Performance summaries of their graduates provided to credential programs should convey the depth, breadth, and appropriateness of each teacher's strategies. However, interpretation of results may be problematic. Teacher performances are likely to be affected both by the choice of topic and by whether or not they have taught the topic, which may not be known to the credential program which prepared them. Guarantees that each teacher has had experience teaching the topic or topics which serves as the focus of the interview would improve the ability to draw inferences about patterns of strengths and weaknesses among program graduates. Since the major focus of semi-structured interviews is on subject-specific pedagogy, any weaknesses in subject-matter knowledge should become apparent. Such weaknesses were identified in many of the beginning teachers participating in the pilot tests.

Questions akin to those in semi-structured interviews are sometimes employed by districts to screen candidates. The utility of a statewide administration beyond a certification of meeting minimum standards is limited. The semi-structured interview's strength is in indicating the depth of a teacher's knowledge in the areas of subject-specific pedagogy and perhaps knowledge of students; however, a district is typically more interested in a teacher's ability to apply that knowledge. The relationship between the depth of a teacher's knowledge and the ability to apply that knowledge is not known, but is likely to be a partial overlap, at best. Districts who are particularly interested in a teacher's ability to teach in a specialized context (e.g., LEP students) are likely to find interpretation of test results beyond pass/fail information problematic.

Semi-structured interviews offer some promise for use in retention, as they focus on an area, subject-specific pedagogy, which was found to be relatively weakly examined in a detailed study of a sample of districts across the State. However, the utility of semi-structured interviews for retention decisions would depend on the relevance of the topic(s) of the interview to the topics which the teacher is currently teaching. This would necessitate continued development of the assessment approach to accommodate differing topics, perhaps based on materials used by individual teachers in their classroom.

For use in professional development, the semi-structured interview format can be very fruitful for use by support providers in evaluating a teacher's skills in subject-matter pedagogy and knowledge of students. However, sufficient time must be available, as semi-structured interviews are very labor-intensive. Standardized semi-structured interviews administered on a statewide basis can provide some diagnostic information, but a teacher's skills are likely to vary considerably with the degree of familiarity with and experience teaching the topic which serves as the focus of the interview. Therefore, results are likely to be topic-dependent; if the topic(s) which served as the focus of the assessment is not reflected in a teacher's current assignment or if the student population is sufficiently unique, then the assessment results may not accurately predict job performance.

Structured Simulation Tasks

Structured simulation tasks address the generalizability problem of other assessments by utilizing a variety of topics and student types in construction of the set of tasks. Performance on each task is affected by the focal topic and type of students, but a sufficient number of tasks is given to enable an assessment of overall general competence in teaching a particular subject. Structured simulation tasks are designed to be relatively inexpensive to administer, compared to other assessment approaches, though not as inexpensive as multiple-choice examinations. Thus, they are well-suited for use in licensure decisions.

Because too few data points are available for subscores, no information beyond a pass/fail score is available for feedback to teacher preparation programs. Therefore, programs whose graduates experienced problems with the assessment would be at a loss to identify the skills that needed to be further developed.

Because structured simulation tasks only provide a measure of generalized competence, they provide little guidance for contextualized hiring and retention decisions by districts. Although it would be more useful for districts to develop their own version of tasks, the developmental costs would be prohibitive.

The lack of indication of strengths or weaknesses also makes structured simulation tasks unsuitable for guiding the choice of professional development activities or for diagnosing an individual teacher's strengths and weaknesses.

Performance-based Assessment Center Exercises

Performance-based assessment center tasks are designed to focus on several distinct teaching skills that can be simulated in assessment center settings. Teaching skills do not lend themselves equally well to simulation, however, with such skills as establishment of rapport, classroom management, and adapting instruction to a particular classroom of students (as opposed to an individual student) being particularly difficult to simulate. If the skills judged as important for assessment purposes lend themselves to simulation, however, and a methodology for evaluating that skill is available, then performance-based assessment center exercises are appropriate for a licensure decision. However, this assessment approach is likely to be one of the more expensive ones because of the complexity of administration and scoring described in an earlier section. As with many other assessment approaches, it is uncertain that the results are generalizable across topics and/or types of students.

Since each exercise focusses on a different teaching skill, some depth of information can be collected on a small number of skills. Information beyond pass/fail statistics could be especially helpful feedback to teacher preparation programs, if the quality of the simulation is high.

Because of the expense of the assessment methodology, performance-based assessment center exercises probably do not lend themselves to local use. Statewide-administered standardized exercises might provide information that could be useful in hiring decisions, although districts would need to know the generalizability of results to the particular context of the teaching vacancy. Performance-based assessment center exercises are less useful for retention decisions, which focus on a teacher's performance in a specific context with a specific group of students.

If the focal skills are of high interest to districts, results of standardized performance-based assessment center tasks can contribute to diagnosis of strengths and weaknesses to guide professional development. Because each exercise typically focusses on a single skill, some depth of information is available which might pinpoint specific areas of strength and weakness within a skill area. However, as with most other assessment approaches, the generalizability of this information beyond the topic and type of students specified in the exercise is unknown.

Videotaped Teaching Episodes

Without major redevelopment, especially in scoring, videotaped teaching episodes cannot fulfill any of the purposes.

Multiple-choice Examinations

The strength of multiple-choice examinations is in the breadth of knowledge assessed, though at the expense of depth, and in the relatively low cost of administration and scoring. They are currently used in licensure decisions as one means of satisfying the subject-matter knowledge requirement, and are in the process of being supplemented by other assessment approaches which allow some measurement of application and/or depth of knowledge. Because teaching is rarely characterized by small, independent decisions which are clearly right and wrong, multiple-choice examinations are less appropriate for measuring general and subject-specific pedagogical skills. This severely limits their utility for feedback to teacher preparation programs, hiring and retention decisions by districts, and professional development.

CONCLUSIONS

This report summarizes two-and-a-half years of research on differing approaches to assess beginning teachers. The evidence to date suggests that there is no one "best" assessment approach. The choice of the optimal assessment approach depends on the skills to be assessed and the purpose for which the information is to be used. If a broad examination of teaching competence is desired, then a combination of approaches is necessary. The preceding sections contain information to guide policymakers in the selection of appropriate assessment approaches.

In addition to considering the information presented regarding each assessment approach, it is imperative that policymakers consider several other factors when selecting appropriate assessment approaches. First, and most important, policymakers must be clear on what they expect a beginning, as opposed to a seasoned, teacher to know and the means by which a beginning teacher is expected to acquire this knowledge (e.g., through teacher preparation, district-sponsored support programs, or thoughtful reflection on experience.) Further, they must have some assurance that adequate opportunities to acquire teaching competencies routinely occur.

Policymakers must also be clear on their relative preferences for testing the depth and breadth of a teacher's knowledge vs. the ability to apply that knowledge in a classroom. As described in this report, no assessment approach examined seems to assess both areas well, though a portfolio with reflective components probably has the greatest potential for addressing both areas. If the interest is in the ability to apply knowledge, then classroom-based assessment approaches are preferable to simulation or other approaches. Simulation and other approaches have difficulty simulating learners, and incorporating the fact that teaching is interactive and highly interdependent with the type of learner. Even assessment materials which include a one-page description of a student fail to communicate important information essential to a teacher's instructional decisions. Also, it is impossible for a simulation to assess a teacher's interaction with students to obtain information and test hypotheses about effective instructional techniques in a particular classroom.

Policymakers should also be aware that for virtually all approaches to assessment, there is a generalizability problem resulting in questionable validity of broad inferences about teaching skills made from a limited sample of topics and teaching contexts. Patterns of effective teaching behaviors vary with subject matter (Stodolsky, 1989), and they are likely to vary also with the type of thinking (e.g., recognition of new concepts, comparison, analysis, evaluation) which is the object of the lesson. Both Stodolsky's work and patterns of beginning teacher performance observed suggest that a teacher's skills vary with differences in subjects (or topics within subjects), types of student grouping patterns, lesson objectives, and teaching contexts. The degree of variance of skills has not been established, and the extent to which skills transfer to different topics, instructional objectives, instructional groupings, and teaching contexts is unknown at this time. Unfortunately, the two assessment approaches that have the potential for overcoming the generalizability problem, structured simulation tasks and multiple-choice examinations, also have severe disadvantages. Multiple-choice examinations are ill-suited to measuring pedagogical skills, and structured simulation tasks cannot provide the diagnostic feedback that would be useful for beginning teachers and their support providers.

While teaching diverse students is an important skill for California teachers to have, policymakers should be aware that none of the piloted assessment instruments did well at assessing a beginning teacher's knowledge and skills in this area. Designing assessments to do so is a formidable challenge. Effective teaching techniques seem to vary with student characteristics (e.g., achievement level or proficiency in English), and no succinct codification of principles across characteristics has been developed. This is especially true for students of differing cultural backgrounds, for example African-American students vs. Chinese-American students. In fact, differences in factors such as social class, fluency in English, and recency of immigration make generalizations even within the same cultural group difficult. Prospective teachers must learn multiple effective teaching principles specific to particular student characteristics. In California, beginning teachers are likely to face extremely heterogeneous classrooms where it is virtually impossible to simultaneously implement all relevant effective instructional techniques, and there is little guidance from the professional literature on how to diagnose and prioritize student needs in these classrooms.

Although the focus of this report has been on the assessment of beginning teachers, the ultimate purpose of the California New Teacher Project (CNTP) is the improvement of teaching in California public schools. The strategy employed by the CNTP for accomplishing this goal is improved assessment and support of the large number of beginning teachers who are expected to enter the profession to both accommodate a growing student population and to replace the large cohort of teachers who will soon reach retirement age. The assessment component of the CNTP has produced a wealth of information about the potential of different approaches to the assessment of beginning teachers. Previous technical reports (Estes et al, 1990; 1992) describe the pilot testing of individual instruments in great detail. This report summarizes conclusions in those reports and makes explicit comparisons between assessment approaches. Analysis of the assessment instruments and beginning teacher performances confirms that teaching is an extremely complex activity; the assessment process and resulting decisions about beginning teachers are likely to be correspondingly complex. Improving the support and assessment of beginning teachers in a cost-effective manner involves many policy choices and tradeoffs which are beyond the scope of this report. However, once such decisions are made, the information contained within this report should assist in the identification of appropriate assessment approaches.

BIBLIOGRAPHY

Ball, Deborah. (1990). *The mathematical understandings that prospective teachers bring to teacher education.* **The Elementary School Journal,** 90,4, 449-466.

Ball, Deborah, and Wilson, Suzanne. (1990). **Knowing the subject and learning to teach it: Examining assumptions about becoming a mathematics teacher. Research report 90-7.** National Center for Research on Teacher Education, Michigan State University.

Berger, J., Connor, J., and McKeown, W. (1974). *Evaluations and the formation and maintenance of performance expectations.* In J. Berger, T. Connor, & H. Fisek (Eds.), **Expectation states theory: A theoretical research program.** Cambridge, MA: Winthrop Publications.

Berliner, David. (1989) *Implications of studies of expertise in pedagogy for teacher education and evaluation* in **New directions for teacher assessment,** invitational conference proceeding. Princeton, NJ: Educational Testing Service.

Delpit, Lisa. (1988). *The silenced dialogue: Power and pedagogy in educating other peoples' children.* **Harvard Educational Review,** 58 (3), 280-298.

Dianda, Marcella, R., Quartz, Karen, J., Radio, Joni, L., and Ward, Beatrice, A. (1990). **Independent evaluation of the California New Teacher Project: 1988-89 report.** Project Document No. SWRL-CNTP-89/90-04. Los Alamitos, CA: Southwest Regional Laboratory.

Dianda, Marcella, R., Ward, Beatrice, A., Quartz, Karen, H., Tusnet, Naida, C., Radio, Joni, L., and Bailey, Jerry, D. (1991). **Independent evaluation of the California New Teacher Project: 1989-90 report.** Project Document No. SWRL-CNTP-90-91-07. Los Alamitos, CA: Southwest Regional Laboratory.

Estes, Gary, Stansbury, Kendyll and Long, Claudia. (1990). **Assessment component of the California New Teacher Project: First year report.** San Francisco: Far West Laboratory.

Estes, Gary, Stansbury, Kendyll, Long, Claudia, and Wolf, Kenneth. (1992). **Assessment component of the California New Teacher Project: Second year report.** San Francisco: Far West Laboratory.

Foster, M. *(1989). It's cookin' now: A performance analysis of the speech events of a black teacher in an urban community college.* **Language in Society,** 18 (1), 1-29.

Goodlad, John. (1990). **Teachers for our nation's schools.** San Francisco: Jossey-Bass Publishers.

Hollins, Etta. (1982). *The Marva Collins story revisited.* **Journal of Teacher Education,** 33 (1), 37-40.

Izu, Jo Ann, Long, Claudia, Stansbury, Kendyll, and Tierney, Dennis. (1992). **Assessment component of the California New Teacher Project: Evaluation of existing teacher assessment practices.** San Francisco: Far West Laboratory.

Ladson Billings, Gloria. (1990). *Culturally relevant teaching.* **The College Board Review,** 155, 20-25.

Lampert, Magdalene. (1985). *How do teachers manage to teach? Perspectives on problems in practice.* **Harvard Educational Review,** 55, 2, 178-194.

Lampert, Magdalene. (1988). *What can research on teacher education tell us about improving quality in mathematics education?* **Teaching and Teacher Education,** 4, 2, 157-170.

Leinhardt, Gaea. (1983). *Novice and expert knowledge of individual students' achievement.* **Educational Psychologist,** 18, 3, 165-179.

Majetic, Richard. (1990). **California Basic Educational Skills Test: Analysis of 1985 examinees and those who repeated the examination in subsequent years.** Sacramento: Commission on Teacher Credentialing.

McDiarmid, G. Williamson, Ball, Deborah, and Anderson, Charles. (1989). *Why staying one chapter ahead doesn't really work: Subject-specific pedagogy.* In M. Reynolds (Ed.), **Knowledge base for the beginning teacher** (pp. 193-205). Oxford, UK: Pergamon.

McDiarmid, G. Williamson, and Wilson, Suzanne. (1991). *An exploration of the subject matter knowledge of alternate route teachers: Can we assume they know their subject?* **Journal of Teacher Education,** 42, 2, 93-103.

Nelson-Barber, Sharon, and Meier, Terry. (1990). *Multicultural context a key factor in teaching.* **Academic Connections, Spring,** 1-11. A newsletter of the Office of Academic Affairs, the College Board, New York City, New York.

Stodolsky, Susan. (1988). **The subject matters.** Chicago: University of Chicago Press.

Ward, Beatrice, A., Dianda, Marcella, R., and Van Broekhuizen, David. (1992). **Independent evaluation of the California New Teacher Project: 1990-91 report.** Los Alamitos, CA: Southwest Regional Laboratory.

Watkins, Richard. (1985). **Third year passing rates on the California Basic Educational Skills Test (CBEST) and passing rates by institution attended.** Sacramento: Commission on Teacher Credentialing.

Veenman, Simon. (1984). *Perceived problems of beginning teachers.* **Review of Educational Research,** 54, 143-178.