ED 336 264 SE 052 155

AUTHOR Collins, Angelo
TITLE Performance-Based Assessment of Biology Teachers: Promises and Pitfalls.
SPONS AGENCY Carnegie Corp. of New York, N.Y.
PUB DATE Apr 91
NOTE 28p.; Paper presented at the Annual Meeting of the National Association for Research in Science Teaching (Lake Geneva, WI, April 7-10, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Biology; Competence; Evaluation Methods; *Portfolios (Background Materials); Science Education; *Science Teachers; Secondary Education; *Secondary School Science; Secondary School Teachers; *Simulation; Teacher Attitudes; *Teacher Evaluation; Teaching Methods
IDENTIFIERS *Performance Based Evaluation; *Teacher Assessment Project

ABSTRACT

Research of the biology component of the Teacher Assessment Project (BioTAP) was conducted to explore the feasibility of using performance-based assessments to evaluate high school biology teachers. Three modes of performance-based assessment were employed: portfolios, portfolio-based simulations, and simulation exercises. Fifteen high school biology teachers completed 11 assessment activities during the 1988-89 academic year and the summer of 1989. Using a holistic scoring procedure and group deliberation, the performances of the teachers were rated. This paper reviews the design and administration of the assessment activities, reports on the rating process, the results of the rating, and the teachers' and research teams' reactions to performance-based assessments. The personnel, portfolio, the assessment center, rating, findings, conclusions, and implications are topics of discussion. A copy of the rating form and 24 references are appended. (KR)

Performance -Based Assessment of Biology Teachers: Promises and Pitfalls

Angelo Collins

Rutgers University

Draft of a paper presented at the Annual Meeting of the
National Association for Research in Science Teaching
Fontana, Wisconsin, April, 1991

The research reported in this paper was supported by grants from the Carnegie Corporation of New York. The opinions expressed are those of the author and not the Carnegie Corpoartion.

Performance -Based Assessment of Biology Teachers:
Promises and Pitfalls

Two of the many aspects of reform in science education are the concern for quality science teaching and concern for authentic assessment. A sign of the first is the summary by the Alliance for Undergraduate Education (1990) of 13 reports on science education published between 1983 and 1989. Although many of the reports focus on the science learning of students, the knowledge and skills of science teachers is of equal concern. This concern about science teaching is a part of a concurrent climate of educational reform, exemplified by the presidential commission report, A Nation at Risk (1983) and the responses to this report. For example, the Carnegie Corporation of New York funded the Forum on Education and the Economy that produced a report, A Nation Prepared (1986), which recommended, among other things, the formation of a national board for and by teachers to set and maintain high standards of excellence for the teaching profession. The National Board for Professional Teaching Standards (1989) is now a reality.

The second concern, arising from growing dissatisfaction with existing modes of assessment that do not have high face validity has launched a movement toward other modes of assessment. Whether termed authentic assessment (Mitchell, 1989; Wiggins, 1989) or performance-based assessment, these assessments hold in common the intention that the assessment will have a high fidelity to whatever is being assessed (Shulman, 1987, 1988). Performance-based assessments are being used for student assessment in art (Harvard Project Zero, 1989), for preservice teacher assessment (Nelson-Barber & Mitchell, 1990) for inservice teacher assessment (Estes, Stansbury, & Long, 1990; National Board for Professional Teaching Standards, 1989; Pecheone, Baron, Forgione, & Abels, 1988), and even for program evaluation.

Intending to inform the deliberations on competent teaching and on performance-based assessment of the National Board for Professional Teaching Standards, The Teacher Assessment Project (TAP) spent from 1986 to 1990 exploring alternative modes of teacher assessment. It was assumed that the research of TAP would also interest teachers, researchers and policy makers concerned with excellence in teaching. Four assumptions guided the research of TAP. The first was that teaching is a complex task and, therefore, the assessment of teachers will require a battery of assessments, some of which will be as complex as teaching itself. No one mode of assessment can be sufficient to capture all the facets of teaching. A second assumption was that teaching takes place in a context -- teaching something to someone somewhere at some time. Therefore, the assessment procedures that were developed were subject specific -- teaching fractions to fifth graders, the American Revolution as taught in eleventh grade social studies, teaching literacy in third and fourth grade, and teaching introductory high school biology. BioTAP, the biology component of the Project, is reported here. A third assumption was that professional teachers have a store of theoretical as well as practical knowledge that supports the decisions they make. The assessments developed were intended to assist teachers in uncovering and explicating their tacit knowledge. A fourth assumption was that the persons who best understand and are qualified to evaluate teaching are teachers. Therefore, all of the work of BioTAP involved active participation of teachers in the design and implementation of the research.

The research of BioTAP was to explore the feasibility of using performance-based assessments to evaluate high school biology teachers. Three modes of performance-based assessment were employed: portfolios, portfolio-based simulations, and simulation exercises. Fifteen high school biology teachers completed eleven assessment activities during the 1988-89 academic year and the summer of 1989. Using a holistic scoring procedure and group deliberation, the performances of the teachers were rated. This paper will review the design and administration of the assessment activities, report on the rating process, the results of the rating, and the teachers' and the research teams' reactions to performance-based assessments. The paper will close with recommendations to those wishing to pursue performance-based assessment of experienced teachers.

Personnel

The Project development team[1], which was responsible for the design and administration and rating of the performance-based assessments consisted of the project director, who had experience in teaching high school biology and in research in science education, four university-based research assistants, three of whom had experience as high school biology teachers, and four Bay Area high school biology teachers. These teachers were nominated by a district or local supervisor, were observed teaching, and were interviewed about their beliefs and practices as biology teachers. Each of the teachers had more than fifteen years experience. They work in schools that represent communities with diverse socio-economic and ethnic populations. All of the development team members believe that teaching and learning science is more than the mere ability to repeat countless, trivial facts. They believe biology is the process of constructing knowledge to explain and predict phenomena about living systems.

In addition to the development team teachers, twenty high school biology teachers were colleagues in the research as they assumed the role of candidate for national recognition, developed portfolios and completed simulation exercises and shared, through interviews and debriefings, the excitement and frustrations of performance-based assessment. (Only fifteen teachers completed the entire set of assessments.) These teachers were selected to represent a variety of teaching contexts and many different years of experience, ranging from 26-year veteran to an intern. Although it was not the goal to select only excellent teachers, this factor did influence the decision about teacher participants. As the teachers on the development team were in regular contact with the teachers doing the assessment activities, the development team teachers were hesitant to commit themselves to working with teachers who had little promise. Lastly, there was an review panel of teachers, science educators, and biologists, who critiqued the research while it was in process and participated, with the development team teachers, in the rating of the completed assessments.

[1] The development team included Angelo Collins, Tom Bird, Ron September, Bruce King, Doug Wong, Susie Turner, Stan Ogren and Gene Gallock.

The Portfolio

One early objective of BioTAP development team was to identify and explore those aspects of teacher knowledge and practice that could best or only properly be assessed by documenting in a portfolio the biology teacher's work in a school or classroom. Evidence of change and growth and of responsiveness to the context were identified as aspects of teaching ideally suited to on-site documentation. A portfolio was defined as a collection of documents that provide evidence of someone's knowledge, skill, and/or dispositions.

Models

When the BioTAP research began, there were few models of a schoolteacher's portfolio. However, other professions present their credentials to members of the profession and to the public by means of portfolios, so it was possible to draw from these images as the concept of a teacher's portfolio developed (Bird, 1990a). An artist's portfolio is a collection of samples of finished, best work that artists agree provide evidence of the knowledge and skill of the artist. The actual samples in the portfolio are interchangeable depending on the goal. The number and variety of documents included in the artist's portfolio is significant -- having too few documents and having too many is equally reprehensible. Too few signals lack of productivity and experience, too many an inability to select quality work appropriate to the goal. As a portfolio, a pilot's log is an ongoing record of work in progress, with commentary, not just records of best or typical work. The catalog of a salesperson indicates that the person has access to and can deliver a variety of materials. The badges of a boy or girl scout indicate an accomplishment that has been achieved with the help of a mentor. The badge has much meaning to other scouts and is awarded with great ceremony and celebration. The intention of BioTAP was to design a portfolio development process for the assessment of teachers that was eclectic, drawing on elements of each of the existing models.

Documents

Another early BioTAP objective was to clarify what the nature of a document might be. There are four classes of documents that a high school biology teacher might prepare as evidence of knowledge and skill of teaching and include in a container called a portfolio.

One class of documents is artifacts, actual samples of the usual work of the teacher. These might include lesson plans or notes about a lesson, sample laboratory instructions or sample tests, or letters to parents or administrators. A special group of artifacts is samples of student work. A second class of evidence is reproductions -- examples of work typically produced in teaching that have no permanence and therefore consciously must be captured in some permanent form for inclusion as a document in a portfolio. Reproductions might include a photograph of a bulletin board or chalkboard display, a xerox copy of student notes, or a videotape of a teacher conducting a lesson. A third class of documents is attestations, reports of the teacher's practice prepared by someone other than the teacher. A letter of commendation by an administrator or a parent or a note written by a colleague commenting on a collaborative project or a letter from a former student are examples of attestations. A fourth class of documents is productions, evidence prepared especially for the purpose of documenting knowledge and skills in a portfolio. A journal, a written reflection, or a document caption are samples of productions. The document caption was recognized as one essential component of an education portfolio, and one of the characteristics that separates the portfolio from other collections of materials such as a scrapbook. The caption is a title sheet attached to each document stating what the document is -- a copy of an overhead showing the interrelationship of the female hormones; what it is evidence of -- an attempt to make a complex, abstract concept concrete; and why it is valuable evidence -- high school science teachers easily become enticed by the words of science and forget how strange these words are to students. Captions for documents were important in the portfolio development process. Documents without captions were meaningless to the raters. Moreover, teachers reported the value of the caption in clarify their intentions and the representation of their practice of teaching.

Design

With these clarification, the research question that was asked to guide the exploration of the development of portfolios by high school biology teachers became: Is a portfolio a feasible mode of teacher assessment? This question implied two sub questions: 1) is it possible for a teacher to construct a portfolio; and 2) is it possible to make warranted inferences from the documents in the portfolio? In addition, the staff wanted the portfolio development process to be an occasion of professional growth for the candidate teachers.

With initial goals and definitions in place, the next task of the development team was to structure the biology teacher's portfolio. The team identified four areas as critical to biology teaching: planning and preparation, instruction, evaluation and reflection, and exchange, which included professional growth and school/community service. These areas provided the structure of the portfolio. In each area, the team then identified many possible activities. In the planning and preparation area, activities included designing the course, planning a unit, planning a lesson, or ordering supplies. The list of possible activities in the instruction area included teaching a laboratory lesson, using materials that supplement the textbook, doing an appropriate and elegant lecture, using co-operative small groups, going on a field trip, and doing a demonstration. The list of activities in the evaluation and reflection area included preparing test items, grading assignments, using a balanced variety of methods for the evaluation of student knowledge and skills, and reflecting on the success or failure of a lesson. The activities in the final area included all of the activities of a teacher that are not included in the other three areas. These might include moderating a science club, making a presentation to the local Audubon Society, consulting a reading specialist or guidance counselor, having a meeting with a parent, attending the annual meeting of the state science teachers association or serving on the school committee to redesign the building. These four areas became the focus of the four entries in the portfolio. An entry was defined as a collection of documents that provide evidence of the knowledge and skill of the teacher about a specific critical activity important to high school biology teaching. For some entries the forms of documentary evidence were prescribed (for example, the instruction entry required a videotape); for others, the choice of documents was left to the discretion of the candidate.

As the teachers on the development team were adamant in reminding everyone, the teachers developing a portfolio for the BioTAP research project were doing this in addition to full teaching responsibilities. With this thought firmly in mind, the decision was made that, although eight entries would be designed, each teacher candidate developing a portfolio would complete five entries in a portfolio -- one in each area and an introduction called background information.

The Entries

Background Information. The first entry which was required of all candidates was Background Information. This entry was not used to evaluate the knowledge, skills, and performance of the candidate, but, based on the assumption that teaching occurs in a context, was to help understand the context in which the teaching took place. This entry provided the candidates with an opportunity to present evidence about themselves. The entry had three parts: 1) a professional biography which inclu.led documents selected by the candidates on their formal and informal education and work experiences; 2) the school and the community setting in which evidence of the socio-economic and ethnic features of the school community were presented by the candidates; and 3) the school environment which included information about the school, the classroom, the students and the teaching responsibilities of the candidates. While some of the evidence for this entry was prescribed, including a required questionnaire, other evidence was left to the candidates' discretion. The candidates were encouraged to include artifacts -- existing school forms and brochures that contained the required information. Only one teacher reported a great benefit from doing this entry. He was consoled as he reflected on the many things he had done because, "sometimes in the classroom it doesn't seem as if I am accomplishing very much." The raters used the Background Information much less than the development team had hoped. Rather than examining this entry first to establish a context, they referred to only it when they had a question. However, each of the other entries provided rich information on context

Unit Planning. In the area of planning and preparation, the team developed an entry on Unit Planning (UE) which all candidate teachers completed. The team members agreed that unit planning was an important and common activity of teaching high school biology and an entry about it would provide evidence of the knowledge and skills of a biology teacher. In the first version of the instructions to the candidate on how to develop this entry, the directions were purposely vague and encouraged teachers to use existing documents.

Make the directions "vaguely prescriptive" was intentional. The team did not want to prescribe the portfolio development process so tightly that some form of teaching that was outside our experience would automatically be eliminated. However, after three months, when about half of the teachers had completed this entry, it was redesigned to be more faithful to the intent of capturing growth and change and to the sequence of planning and teaching. The second version, termed the construction kit, was more prescriptive Teachers were specifically asked to document where the unit was located in the course, what resources they intended to use, to complete a lesson log sheet for each day of the unit including the intentions (what will the student learn today), the reasoning (how this lesson helps achieve the purposes of the unit), the assessment (how did the lesson go, compared with your intentions), and the adjustments (in light of what happened, how will you adjust your plans), and and to write a reflection after the unit was completed. Although each candidate completed only one planning entry, they all completed some original versions and some construction kit versions of entries. For the most part, they preferred the construction kit entries because they knew exactly what to do and what was expected of them. When it came to rating the portfolios, there were no differences based on the entry version used. That is, all construction kit entries were not rated higher than all original version entries. Recall that these teachers in the BioTAP research were developing portfolio without models. As portfolios become more common in education, the question of the degree of prescription in the directions will need to be revisited, based, in part, on the models of portfolios these teachers created.

Instruction. Two entries were designed for the instruction area -- teaching a lesson Using Alternative Materials (ME) and teaching a Laboratory Lesson (LE) -- but each candidate only submitted one. There was no disagreement among the members of the development team that an entry on instruction was necessary to provide the opportunity for teachers to demonstrate their knowledge and skills. The laboratory lesson was chosen because the team believed that safe, effective laboratory experiences are of the essence of biology teaching and learning. A laboratory lesson was defined as one consisting of "hands-on, minds-on activity involving the manipulation of living or non-living materials, equipment and/or data. During this lesson the students will engage in some scientific thinking such as reasoning, analysis, prediction, hypothesis formation or evaluation (Collins, Bird, King, & Sel .ember, 1988, p. 22)." The alternative materials entry was chosen because of the belief that the textbook is not the curriculum and therefore, teachers must supplement the textbook to enhance, simplify, update, or add context to instruction. In both instruction entries, the candidate had to present a videotape of the lesson. In addition to the videotapes, the candidate indicated what instruction preceded and what followed the videotaped lesson, included samples of student work associated with the lesson, and wrote a reflection on the purpose of the lesson and how he or she determined if the lesson was successful. As in the unit planning entry, the construction kit imposed order on the collection of documents.

Student Assessment. The entry designed for the evaluation and reflection area focused on Student Assessment (SE) and required the candidate to maintain a journal of all the forms of evaluation for a six week period. Of the many possibilities for an entry in this area, it was agreed that a teacher, adapting evaluation to the context of the students and the topic, employs a variety of methods to evaluate the knowledge and skills of students and that examining a journal of the methods of evaluation would provide evidence of teaching performance. From the journal, the candidate selected four different methods of evaluation that were employed and wrote a detailed analysis of the rationale for, success and failure of that method. The analysis included work samples from three students and responses to probing questions such as "Summarize your view of the student's progress and problems; Describe the feedback that you gave or need to give this students; and Note how this student's performance did or should affect your subsequent instruction (Collins, Bird, King, & September, 1988, p. 35)."

Exchange. The entries in the exchange area also took the form of a journal. The candidate maintained a record of exchanges either with education professionals or with members of the community for a period of six weeks and then commented on two of these exchanges in detail. This area was difficult to define and including it in the portfolio was controversial. The teachers on the development team were adamant that a teacher they would want to be recognized as someone who met high standards of excellence would not only be a well-prepared, caring, elegant instructor, but also would be a good citizen of the school, local and/or professional community. However, they were cautious about recognizing a person who was excellent at giving workshops and attending committee meetings but was not good in the classroom with students. Members of the advisory committee who critiqued the work of BioTAP as it was in progress were not unanimous in their support of this area. Some felt it was unjust to assess a teacher on anything other than the act of teaching. However, one of the five core propositions of what a teacher should know and be able to do proposed by the National Board for Professional Teaching Standards is that the teacher is a member of a learning community. Designing an entry for the area was also difficult. In retrospect we recognized that, unlike the other entries where we identified instances of the critical task area and selected one or two, in the entries in this area we tried to provide opportunities to document just about any non-instructional exchange. The defined entries of Professional Exchange and Community Exchange were not useful divisions. They did not provide mutually exclusive domains of activity to document. Neither were they inclusive of all possible types of non-instructional exchange. The requirement to have others attest that the exchange actually occurred was problematic for some candidates. Like the background information, these entries were interesting to review, and several teachers reported that keeping the journal made them aware of how much they were doing and forced them to consider re-evaluating their commitments. However, these entries were not critical in the rating the performance of the teachers.

The assumptions, intentions, definitions and directions for the portfolio development process for high school biology teachers were compiled into a Handbook for Development of the Biology Teacher's Portfolio (Collins, Bird, King, September, 1988). In Fall, 1988, the Handbook was presented and explained to twenty teachers at an orientation meeting. Throughout the year as the biology teachers developed their portfolios, the development team members called them, visited them at their schools and invited them to come together for support and sharing. Based on their individual timelines, each teacher sent in portfolio entries as they were completed. Based on the early returns, a new handbook, called the Construction Kit, which was more prescriptive, was written and given to the teachers in January, 1989. All of the portfolios were completed by mid-May, 1989.

The Assessment Center.

With the portfolio development process launched, the BioTAP Development Team began to consider the design of the Assessment Center to be held in June, 1989 in which simulation exercises and portfolio-based simulation exercises of critical tasks of biology teaching would be administered and rated.. The assessment center would last six days -- two days to rate portfolios, learn to administer exercises, and design portfolio-based exercises; two days to administer simulations; and two days to rate the assessments and reach a decision about the performance of each candidate. The Assessment Center had four goals: 1) to adapt exercises that had been designed earlier for mathematics and history teachers to biology teachers; 2) to design exercises that were based on the portfolio entries; 3) to test a form of holistic rating; and 4) to provide an experience that would be educational for the candidate as well as for assessment.

The reason for the first goal was to discover if it was possible to consider an exercise as if it were a shell or template for any content area and then add detail to make it specific for a given content area. We also wanted to draw on the wisdom and experience from the earlier work of the Project contained in the extensive and detailed technical reports that had been written about using exercises to assess teachers of mathematics and history. The second goal for the assessment center was to design simulation exercises based in portfolios. One reason to do this was to correct for faulty portfolio design. However, the prime reason to attempt portfolio-based simulations was to capitalize on the best aspects of two modes of assessment -- simulation exercises and portfolios. The simulation exercises in first phase the Project had been criticized because they did not capture enough of the context of the teacher in her own classroom. Portfolios were being criticized because they were not standardized and therefore it would be difficult to compare performances. Portfolio-based simulations were to provide the integration of context from the portfolio and standardization from an exercise (Vavrus and Collins, 1989; King, 1990a). The third goal for the assessment center was to use a holistic rating system based on professional judgement rather than doing a fine grained analysis of each performance. Our final goal had as much to do with the conditions of teaching as with assessment. In our discussions we kept returning to the point, "Wasn't it a shame that all these excellent teachers and science educators would be together for a week and never get to share their wisdom and experiences with each other." A formal occasion, an exercise that facilitated professional sharing, would provide such an opportunity.

Design

In designing simulation exercises, with our mission of exploration and a sample of less than 20 candidates to examine, we were not after generalizable results or systematic comparisons among evaluation procedures. A thoughtful array of suggestive cases, examples, and anecdotes was our aim. One consideration was the relationship between the exercise and the portfolio: self-contained exercises with no connection to the portfolio called simulations; exercises situated in the context of portfolio entries called portfolio-based simulations; and simulations designed to compensate for deficiencies in the portfolio design called portfolio extensions. Another consideration was the relationship between the examiner and the candidate: unsupported performance; supported performance; and test-intervene-retest

In addition, we imposed several other constrains on ourselves as we designed the simulation exercises. We wanted to keep the ratio of examiners to candidates to 1:2. This meant that every exercise could not be administered as a one-on-one interview; some forms of group exercise administration would be necessary. In addition, it seemed appropriate that computer technology play a role in the assessment of science teachers.

Added to choices of exercise types and constraints, there were two other types of decisions to consider: how would we present the problems to the candidate, and how would we arrange for the candidate's responses. In presenting the problem to the candidate we looked for opportunities for candidates to be engaged in many different ways. The list of possibilities included: questionnaires, writing tasks, interviews, agendas for discussion, what-if situations, vignettes with questions, role-playing, and things to play with, work with, write about, talk about. In arranging for candidate responses we considered: interactive/immediate response (face-to-face interview), self-paced/delayed response (questionnaires and computer administered), oral, written, enacted (e.g., say it to me the way you would say it to your students), situated or unsituated (telling about something that one is doing, or holding, or looking at), generating a comment on something that is not here and not now. We also needed to clarify what we want to discover about the candidate in the exercise and how to make a hypothetical situation vivid for the candidate.

S'mulation Exercises

With the design elements articulated and the goals and constraints identified, it was decided that we would develop eight exercises, but each candidate would only do seven.

Unit Planning Review. This exercise (UX) was a portfolio extension exercise, intended to probe for a deficiency in the Unit Planning Entry in the portfolio. In reviewing the early portfolio entries on unit planning, there seemed to be, despite specific questions, an emphasis on content and a dearth of information about the role of concern for students in the planning process. Also, many of the critics of the Project had expressed concern that there was not enough emphasis on issues of equity in the assessments, and so this exercise became one opportunity to probe candidates about issues of equity in classrooms with diverse student populations. The form of this exercise was test -- intervention -- retest. In the exercise, the candidates were given an opportunity to review their portfolios before they came to the assessment center (the test), they then participated in a structured discussion of issues of equity and diversity in the classroom (intervention) , and then wrote responses to four questions specifically designed to probe about how the unit plan met the needs of certain students in the class and how the plan would be altered if their were different students in the class (retest). For administration, two examiners participated in discussion groups of four candidates, but the writing was done alone and unsupervised.

The Student Evaluation Exercise. This exercise (SX) was a portfolio-based interview. There were three different activities in the exercise: standard questions, tailored questions, and a role-play. The standard interview questions were written before the assessment center and were answered by every candidate. The tailored questions were written in the first two days of the assessment center by the examiner who reviewed the portfolio and administered the exercise. These questions were specific to the candidate and the evidence in the portfolio. The last part of the exercise was a role-play -- the examiner chose a sample of student work from the portfolio and asked the candidate "What if I were a parent of this student and wanted to know about my son's progress..."

Laboratory Monitoring Situations Exercise. This exercise(LX) was designed as a portfolio-based exercise situated in the Laboratory Lesson Portfolio Entry. The intention was to probe the candidates knowledge about classroom management, time management and student misconceptions. As the exercise, in interview form, was piloted tested with members of the development team and review panel, the pilot test teachers responded to the questions by telling stories about their classes. With the decision to make a virtue out of reality, the exercise was redesigned so that each candidate would tell six stories, two each about three hypothetical problems occurring in the laboratory lesson described in the portfolio entry. Two stories were told to provide candidates opportunities to add context to their stories that would alter their plan of action. The exercise was designed for computer administration and, although candidates had the option of writing responses rather than entering them in the computer, none chose to do so.

Analyzing the Alternative Materials Lesson Exerci This simulation (MX) was an extension of the Alternative Materials Portfolio Entry (ME) and was designed to probe more deeply the understanding that teachers had of what was non-textbook material. BioTAP was disappointed because so many of the portfolio entries had focused on a lecture as the non-textbook oriented lesson as we had thought we had designed an entry that we would get a variety of lessons -- hands-on exercises, demonstrations, movies, student projects. In retrospect, the portfolio entry directions may have been too vague.

In any case, we were still interested in when, why, and how, teachers stepped outside the textbook material for instructional purposes. The exercise was designed as an interview and administered in one of three forms, an one-on -one interview, as a written questionaire and on the computer.

The Videotape Reflection. This simulation (VX) was modified from exercises for teachers of mathematics and history. The assumption behind the exercise is that, if teachers are going to control membership in their profession, they must be able to recognize good and poor teaching when they see it and offer advice to other teachers. In this exercise, two segments of videotape were used, both of teachers conducting a lesson on a critical issue in biology. One tape segment was of a small class where a student-led discussion on issues about abortion was taking place; the other of a large class with a teacher-led discussion on dangers in recombinant DNA research. The candidate viewed the videotape and answered questions in a one-on-one interview about the teaching and about how he deals with controversial issues in his classroom.

The Coping with a Biology Textbook Exercise. This simulation (TX) was modified from exercises for history and mathematics teachers from the earlier work of the Project. The focus of the exercise was adapting a section from the text using a book that had been assigned by the school district. Much of the emphasis was on the content knowledge of the topic, ecology. The exercise was administered on a computer, although a written version was available for anyone who might select it. No one did.

The Computer as an Instructional Tool Exercise. This simulation (CX) was a modification of an exercise in mathematics on using computers. In the exercise, the candidate was required to design a lesson plan for several gifted but bored students, using the computer program on genetics as as one of the instructional tools. The candidate had time to use the program and sketch their instructional plan before a one-on-one interview.

The Deliberations on a Problem Exercise. This exercise (DX) was administered to a group of four candidates and two examiners and intended to give candidates an opportunity to "bring it all together.". A formal process of turn taking was designed in which each candidate expression an opinion on some question about teaching evolution. Then there was an open discussion, including the examiners. It was the last exercise, administered to all candidates at the same time.

The Biology Assessment Center was held in June, 1989. Fifteen teachers completed all of the portfolio entries and the appropriated combination of simulation exercises.

Rating

At this time, it must be made very clear that the biology teachers who completed portfolios and simulation exercises were truly pioneers. They agreed to participate in the research knowing they would not receive personal feedback or scores on their portfolio performance. It would be impossible for the research to evaluate the modes of assessment and the performance on these modes of assessment at the same time. The performance rating on an untried assessment would be meaningless. Yet, to test the portfolio as a mode of assessment, the performances had to be rated to determine if it was possible to do so. If all the candidates got low scores on the same assessment, that would just as likely indicate something about the assessment as about the teacher.

With the portfolio development started and the simulation exercises being designed, the development team turned its attention to the serious question of how to rate performance-based assessments. Earlier work by the TAP staff had developed data-driven, fine-grained scoring systems for the simulation exercises in mathematics and history. It had taken many months to design a scoring system for each individual exercise and many more months to score each exercise (Kerdeman, 1989). For contrast, BioTAP decided to design a holistic rating scheme, based on professional teacher judgement.

Criteria. For new modes of assessment, such as portfolios and simulations, designing a rating scheme is fraught with problems. For example, if the criteria are determined too early in the process, some forms of excellence might never be seen, because the teachers would match their evidence to the criteria. However, if the criteria are not determined soon enough, the teachers have the right to ask, "What do you want?" The BioTAP staff addressed this problem by hinting at the criteria that had evolved during the research on the assessment of teachers of mathematics and history in a section in the Handbook called justification.

Another issue in designing a rating system is the question of whether or not to use a compensatory model of assessment. Can a teacher be rated so high on planning that his low score on instruction will be canceled? Must a teacher meet a minimum competence on all criteria on all portfolio entries to be judged successful. In scme ways we begged this question, by placing the decision about certification in the hands of a caucus of the raters.

For several months, the BioTAP development team played with different rating schemes using the first portfolio entries that were submitted. Finally, the BioTAP staff to devise a rating scheme derived from the five core propositions of what a teacher should know and be able to do presented by the National Board for Professional Teaching Standards (1989). The rating categories became: 1) The candidate[teacher] attended to students and their learning; 2) The candidate knew the subject matter and how to teach it; 3) The candidate attended to class management and monitoring; 4) The candidate thought about and learned from his/her activity; and 5) The candidate participated in a learning community (Collins, Bird, King, & September, 1989).

The final rating form used for all BioTAP assessment activities is found in Figure 1. The form was designed to encourage the raters to rely on their professional judgement and to allow the research team to trace their decisions (Bird, 1990b).

Insert Figure 1 Here

At the top of the form are spaces for identification and a box in which the rater wrote a brief description of the portfolio entry, "Videotape and student worksheets from a sophomore class on cat dissection." Below the box are the five rating categories and four rating scales: relevance, evidence, difficulty and goodness. Relevance was considered an important scale because the development team anticipated that certain portfolio entries would be more likely to yield evidence for some rating catego.ies than for others. For example, we expected that the portfolio entry on instruction would yield more evidence in the category that "the candidate knew the subject matter and how to teach it" than it would in the category on "the candidate participated in a learning community." This expectation was confirmed during the rating sessions. The scale on evidence was used to allow raters to express their confidence in their judgement based on the amount of evidence that was present in a portfolio entry. For example, for the unit planning entry, one teacher candidate wrote an extensive reflection after each lesson (as she had learned in her teacher education program) while another had a one sentence reflection at the end of the unit. There was a difference in the quantity of evidence in the portfolio entry. The third scale on difficulty was included to allow the rater to place qualifications on the nature of the performance the teacher included in the portfolio. For example, is conducting a laboratory lesson on recombinant DNA more difficult that teaching a laboratory lesson on using a microscope to look at cells? In the rating process, this scale was seldom used. For each of these three scales, the range of scores was one to three. For each category for each performance, a default score of two was placed in the rating box. The rater had to change a default score to three, for example, to indicate that the entry was especially relevant, rich in evidence, or difficult and write an explanation at the bottom of the page about why the rating was changed.

The final scale was the goodness scale. The range of scores was one to five, with an option for a score of zero if a rater felt she could not form an opinion about the performance. A score of one indicated that the rater felt the performance was "unacceptable, even from a novice;" a score of four indicated that the performance was "proficient from an experienced teacher." For the goodness scale, a final category of overall rating was added. At the bottom of the rating form was space for the rater to write notes explaining the scores that were given.

Sixteen people were involved in rating the assessment materials. The raters were high school teachers, most biology, university-based science educators, and research biologists Before they assembled, the raters had an opportunity to read the rating procedure guidelines. Training on the use of the rating form was minimal. The staff member who had done most of the work on designing the rating manual explained how to use it. The raters then used the form to rate his teaching performance, and they asked questions.

The sixteen raters became expert on rating a portfolio entry or a simulation, rather than on a candidate. The raters also administered the simulation activities to the teachers. Rating portfolios was completed before the simulations were administered. Rating simulations was completed as soon after the teacher performance as possible. All teaching performances were rated twice, in the case of the simulations, by the assessment administrator and by someone relying on notes and/or an audiotape. After all performances were rated, the rating sheets were collected, duplicated, and collated. A rating packet was constructed of all the rating sheets for all the activities for each teacher candidate. Each rater was then given the rating packets of five teachers and asked to read them and make an overall decision about the performance of the teacher. The raters then met in caucus groups of four, who had reviewed the packets of the same five teachers. The caucus groups then reached a consensus about whether or not the teacher would be "certified," acknowledge as eligible for a national certificate of excellence. Later all the ratings were averaged and numerical scores attached to each teachers performance and to each raters score assignment

Findings

The findings can be grouped into five categories: 1) feasibility; 2) differences in ratings; 3) differences in performance; 4) teachers reactions; and 5) design team reactions.

As trite as it may seem, the primary result is that teachers can do performance-based assessments and these activities can be rated and they do discriminate.

Differences in Ratings

During the caucus discussions of the teachers' performances, only one teacher was deemed "certifiable." However, three other teachers were placed very close to certification. Two of these teachers were in their first two years of teaching and it was suggested by the raters that they each teach another year to provide depth to their responses. The other teacher was advised to repeat one portfolio entry, as it appeared inconsistent with the rest of the assessment activities.

The range of overall averages of the mathematically calculated scores was 2.636 to 3.615 The mathematical scores were not isomorphic with the decisions made by caucus deliberation. The two numerically high scoring teachers who were not accorded high ratings by deliberation had not complete all the assessment activities.

Based on 307 overall ratings, given to the candidates on any exercises that were completed, being rated twice on each, the most frequent rating was 3 (acceptable). The activities for which teachers were most frequently rated unacceptable (1 or 2) were the portfolio entries on student evaluation and unit planning, and the simulations of textbook adaptation and unit planning. The activities most frequently rated as more than acceptable (4 or 5) were the portfolio entries on unit planning and professional exchange, and the simulations on using a computer, deliberating on teaching evolution, and critiquing the video of other teachers. Note that the unit planning portfolio entry had both the most acceptable and most unacceptable ratings. However, the student evaluation portfolio entry had the widest range, being the only assessment with scores in all five scale groups. The simulation based on instruction had the smallest range with all but 4 ratings being average.

Another difference in rating is the judgements made by the different raters (Collins, 1990a) . By looking at the comments on the front of the rating forms, it is possible to infer what the different raters emphasized in the rating process. One high school English teacher with more than 20 years experience rated the assessments. This was done to determine if non-biology teachers could rate the assessments of biology teachers. Most of her rating comments wei e general, and focused on what an English teacher is expected to knew -- students, teaching strategies, reflection and rationale. One research biologist rated biology teacher assessments to determine if a person without school experience could rate teacher performances. The most noticeable characteristics of the comments written by the biologist was their length -- they are more than three times longer than any other set of comments by any other rater. The biologist's comments about subject matter were unanticipated -- never made without a reference to teaching. It had been expected that she would look sharply at how accurate and how current the content knowledge was. It was not surprising that the biologist did not comment on classroom management. The three university science education faculty members were raters The comments made by the science educators were broad, judgmental claims, uniformly terse. Although comments were made by the science educators about the teachers' concern for students and their learning, these comments were less common than from other raters. The science educators were explicit in addressing the teachers knowledge of subject matter and how the teacher taught the subject matter and commented frequently on the teachers reflections and rationales. Nine experienced high school biology teachers constituted the largest group of raters In every set of comments by these teachers there was a phrase about how well the teacher attended to students and their learning. The high school biology teachers placed most emphasis in their rating on the two activities that are most likely to take place in the classroom -- student learning and teaching biology. They placed relatively little emphasis on the teacher assessment activities that do not capture teaching as it is currently practiced -- reflection and participation in learning communities. The experienced biology teachers were the only group of raters to match what teachers said they did and what the evidence indicated they actually did.

Differences in Performance

After all the assessment activities were completed and rated, the materials of the four highest rated teachers and the materials of the four lowest rated teachers were compared. The materials from those teachers that received high ratings differed from the materials from the low rated teachers in that they were: 1) very student centered and 2) clear and explicit, for example about what a document was evidence of. (Aninao, 1990). Collins (1990b) compared the porttfolio entries on planning of all the biology teachers and found that those that were rated as a 3 or higher were organized and easy to follow and all of them included a justification for teaching what they did.

Teacher Reactions

In discussions with the teachers during portfolio development, after the simulation exercises were completed, and a year later, there was unanimity that participation in the research had influenced their practice of teaching for the better. They felt that many of the activities had a high fidelity to their practice, that they had been challenged and had become more reflective during the time spent on the portfolio development and simulations (King, 1990b). Further, they felt that the student evaluation portfolio entry with its follow-up simulation had the highest fidelity to their practice and that the exchange entry was pointless, besides making them feel good. However, they also stated that the assessments had been hard work and would not be worth the effort without some reward or compensation.

Development Team Reactions

Among the development team it was felt the array of assessments captured most aspects of the complex act of teaching biology. Further, it was believed that the exchange entries had not successfully captured what we had intended, but that another activity to capture how teachers are members of learning communities was needed. In determining the value of the portfolio-based simulations, the ones that were most closely tailored to the portfolio (in student evaluation, more than half the probe questions for each teacher interview were unique to his/her portfolio) and the ones that were least tailored (in the lab instruction follow-up the teachers were asked to tell stories and they often situated their story, not in the lab documented in the portfolio, as directed, but in another lab) were most successful in capturing what teachers knew and did. The discussion on evolution and the use of the computer had potential as valuable assessment activities, but the first became bogged down in technique, and the second contained too many different tasks in too short a time. We also determined that simulation activities can be modified from one content domain to another, but only with extensive revision.

Conclusions and Implications

Although the eleven assessments design by BioTAP are not the only way to dissect the act of teaching biology, they demonstrate that it is possible to design performance-based modes of assessment. Although time consuming to design, develop and evaluate, performance-based assessments can be completed by teachers, can be rated and do discriminate. The process of designing performance-based assessments is a tool for examining assumptions about biology teaching and characteritics of good teaching. The design process provided opportunities for reflection. Among the practical things learned by BioTAP was the value of the caption in attached to the documents in the portfolio.

The teachers who completed the exercises felt that the assessments captured the complexity of teaching. Further, completing the assessments provided teachers with opportunities for reflection and clarification.

The psychometrics of performance-based assessments of teachers is still in its infancy. Questions of validity, reliability, generalizability need to be addressed as the research and implementation of performance-based assessment continues. Heartel (1990) among others, has begun such work.

Much more work needs to be done on the design, development and rating of portfolio, simulations, and still undiscovered modes of performance-based assessment. Further research needs to be done on how these performance-based assessments have functions other than assessment, such as in teacher education programs. New research questions about the effects of such performance-based assessments will need to be addressed. The exploration of performance-based assessments of biology teachers has opened many possibilities for research in science teaching.

References

Alliance for Undergraduate Education (1990). The Freshman year in science and engineering: Old problems, new perspectives for research universities. Washington, DC: The National Science Foundation.

Aninao, J. (1990). The teacher's written plan: Its problems and potentials as an assessment tool. (Technical Report N 5). Stanford, CA: Stanford University, Teacher Assessment Project.

Bird, T. (1990a). The Schoolteacher's portfolio: An Essay on possibilities. In J. Millman & L. Darling-Hammond (Eds.) Handbook of Teacher Evaluation: Elementary and Secondary Personnel, Second Edition, Beverley Hills: Sage. (pp 241-256).

Bird T. (1990b) Report in the rating procedure used to assess portfolios and assessment center exercises for high school biology teachers. (Technical Report B3). Stanford University, Teacher Assessment Project.

Collins, A. (1990a, April). Novices, experts,veterans and masters: The role of content and pedagogical knowledge in evaluating teaching. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Collins, A. (1990b, April). The teacher's portfolio -- What is necessary and sufficient? Paper presented at the annual meeting of the American Educational Research Association, Boston, MA

Collins, A., Bird, T., King, B., & September, R. (1988). Portfolio development handbook for biology teachers. (Working Document, B20) Stanford, CA: Stanford University, Teacher Assessment Project

Collins, A., Bird, T., King, B., & September, R. (1989). Biology examiner's handbook for assessment center. (Working Document, B23) Stanford University, Teacher Assessment Project.

Estes, G.D., Stansbury, K., & Long, C. (1990, March). Assessment component of the California new teacher project: First year report, Technical Report, San Francisco, Far West Laboratory for Educational Research and Development.

Haertel, E., (1990, April). From expert opinions to reliable scores: Psychometrics for judgement-based teacher assessments. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, MA.

Harvard Project Zero. (1989). Portfolio: The Newsletter of Arts PROPEL. 1 (5).

Kerdeman, D. (1989). The 100 statements project: A study in the dynamics of teacher assessment. (Technical Report, E11). Stanford University, Teacher Assessment Project.

King, B (1990a). Thinking about Linking portfolio entries with assessment center exercises: Examples from the Teacher Assessment Project . (Tech. Rep. N2) Stanford, CA: Stanford University, Teacher Assessment Project.

King, B. (1990b). Teachers' views on performance-based assessments. (Tech. Rep. N3) Stanford, CA: Stanford University, Teacher Assessment Project.

National Board for Professional Teaching Standards. (1989) Toward high and rigorous standards for the teaching profession. Washington, D.C., National Board for Professional Teaching Standards.

Nelson-Barber, S. & Mitchell, J. (1990) Variations on a theme: Regional nuances in the assessment of teachers. (Technical Report T1). Stanford University, Teacher Assessment Project.

Mitchell, R. (1989). What is "authentic assessment?" Prepared for "Beyond the Bubble" Curriculum? Assessment Alignment Conferences Cosponsored by the County State Steering Committee of the California Association of County Superintendents of Schools and the California Assessment Program, State Department of Education.

Pecheone, R.L., Baron, J.B., Forgione, P.D., & Abels, S., (1988). A Comprehensive approach to teacher assessment: Examples from math and science. In A. Champagne (Ed.) This Year in School Science: Making the System Work, Washington, D.C. American Association for the Advancement of Science. (pp 191-214).

Shulman, L.S. (1987, May). Assessment for teaching: An Initiative for the profession. Phi Delta Kappan, (pp 38-44).

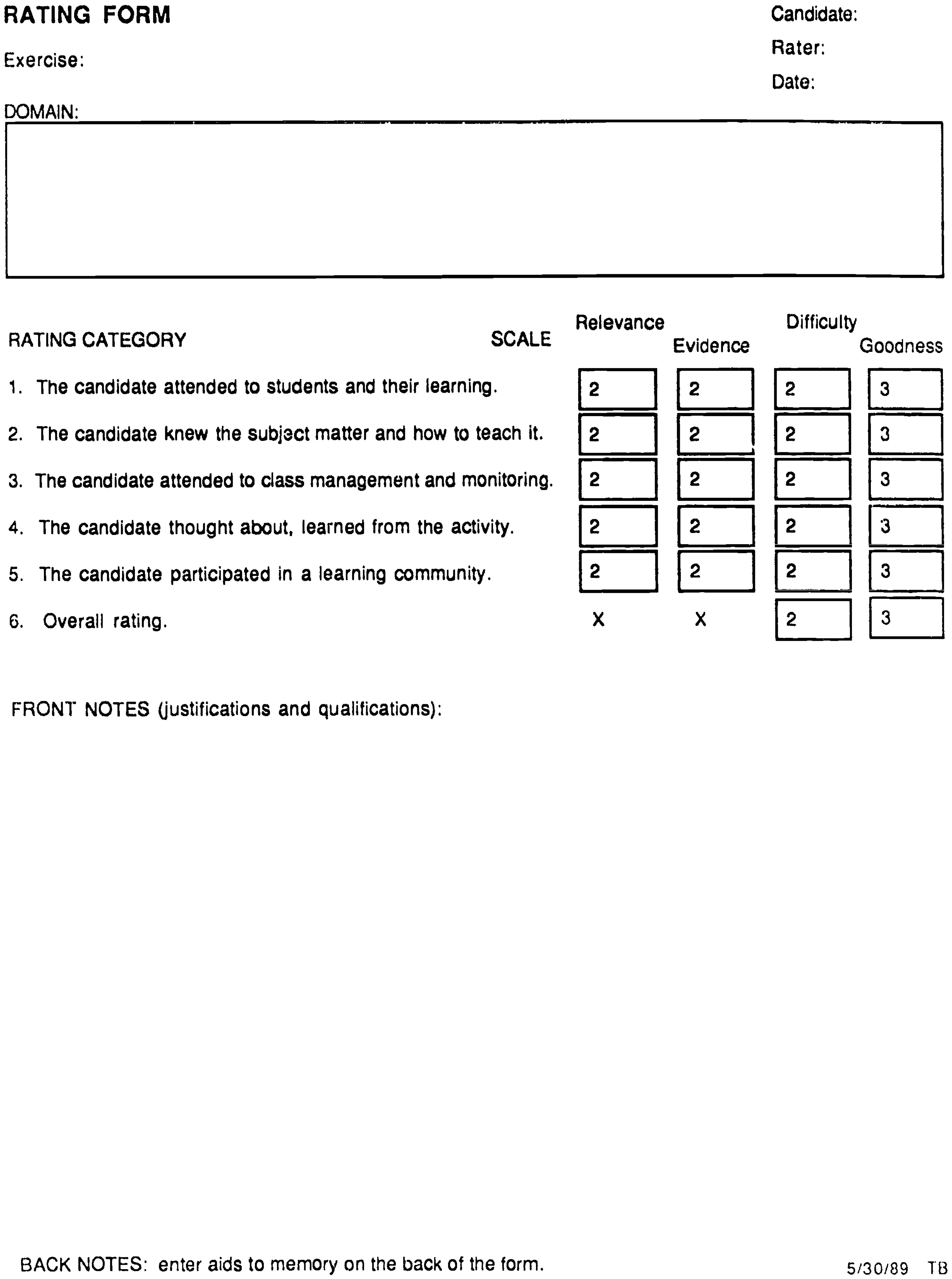
Shulman, L. S. (1988, September). A Union of insufficiencies: Strategies for teacher assessment in a period of educational reform. Educational Leadership.

Vavrus, L. & Collins, A. (1989, April). Portfolio documentation and assessment center exercises: A Marriage made for teacher assessment. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA

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

____ (1983) A nation at risk: A report to the nation and the secretary of education. United States Department of Education by the National Commission on Excellence in Education, U.S. Government Printing Office.

____ (1986) A nation prepared: Teachers for the 21st century. by the Carnegie Forum on Education and the Economy, Carnegie Corporation of New York.



RATING FORM

Exercise:

Candidate:

Rater:

Date:

DOMAIN:

RATING CATEGORY	SCALE			
	Relevance	Evidence	Difficulty	Goodness
1. The candidate attended to students and their learning.	2	2	2	3
2. The candidate knew the subject matter and how to teach it.	2	2	2	3
3. The candidate attended to class management and monitoring.	2	2	2	3
4. The candidate thought about, learned from the activity.	2	2	2	3
5. The candidate participated in a learning community.	2	2	2	3
6. Overall rating.	X	X	2	3

FRONT NOTES (justifications and qualifications):

BACK NOTES: enter aids to memory on the back of the form.

5/30/89 TB

Figure 1. Rating Form