Developing a National Science Assessment for Teacher Certification:

Practical Lessons Learned

Steve Schneider
Kirsten R. Daehler
West Ed

Kristin Hershbell
City College of San Francisco

Jody McCarthy
(formerly of WestEd)

Jerome Shaw
University of California Santa Cruz

Guillermo Solano-Flores
University of Colorado at Boulder

ABSTRACT

INTRODUCTION

Creating something entirely new, something important, something for which there is no agreed upon "right way," set model, or solid precedence is exciting—and, at times, frustrating. Developing the Adolescence and Young Adulthood (AYA) assessment for science teachers of students aged 14 to 18+ for the National Board for Professional Teaching Standards (NBPTS) is a case in point. This chapter describes our experiences as an Assessment Development Laboratory (ADL) and looks at some of the challenges inherent in developing a large-scale assessment that is complex, strives to be innovative and must be closely aligned with a given set of standards. Some of the external challenges we faced included shifting and unclear expectations, the conflicting needs of multiple stakeholders and a deadline that was dramatically shortened midway through the process. Within the assessment development process itself we also needed to consider how best to involve teachers, address issues of equity and standardize the process to maximize efficiency. We share some stories to illustrate not only the challenges but also the insights gained and lessons learned with the hope that they provide a useful historical perspective relevant for other large-scale assessment development projects.

THE WESTED ASSESSMENT DEVELOPMENT LABORATORY

The Assessment Development Laboratory team at WestEd[1] was formed in the spring of 1994 when the NBPTS awarded WestEd a two-year contract to develop the Adolescent and Young Adult/Science (AYA/S) assessment aligned with the *Adolescence and Young Adulthood/Science Standards for National Board Certification* (hereafter referred to as *"AYA/Science Standards*"; please see Appendix A for an overview of the standards). This type of standards-driven assessment for teachers was a natural spin-off from the national trend in standards-driven reform for students, itself an outgrowth of systemic reform in the 1980s (Smith & O'Day, 1991). The NBPTS's competitive bid asked ADLs to develop a two-part assessment system consisting of a portfolio and a series of assessment center exercises. In doing so, ADLs of all content areas were required to undertake the tasks outlined in Figure 1.

<Insert Figure 1 about here>

To meet this challenge, WestEd assembled a diverse team of respected, high caliber teachers, measurement experts, and leading science educators. Steven Schneider, the director of WestEd's science and mathematics program, served as co-Principal Investigator along with Mary Budd Rowe, professor of Science Education at Stanford University. Mary Budd Rowe, an original member of the NBPTS, provided an important link between the WestEd team and the NBPTS. She brought the history of the NBPTS to WestEd and offered wisdom in a familiar face to the NBPTS.

Other core team members and key advisors included past presidents of the National Science Teachers Association, teachers who were also members of the NBPTS's AYA/Science Standards Committee, and representatives from internationally recognized informal science education organizations, including Lawrence Hall of Science and the Exploratorium. Being geographically close and well connected to Stanford University allowed the team to call on the expertise of the NBPTS's visionaries, such as Lee Shulman, and other leading science educators and psychometricians. Most importantly, our team was composed almost entirely of past and present teachers. This made us better able to listen carefully to other teachers, thereby maintaining connections, credibility, and an allegiance to what science teachers value. While the high caliber of WestEd's team members and our close connections to many important stakeholders helped us to work in a highly politicized environment, we were not completely buffered from difficult circumstances and internal tensions.

Assessment Development: Starting Up and Skidding to an Abrupt Halt

When we began our work with the NBPTS in the spring of 1994, it was with a healthy sense of idealism and in a climate supportive of innovation. Helping to fuel our enthusiasm was the knowledge that we were part of a larger effort. After all, a handful of other new ADLs had also just received funding from the NBPTS to develop assessments in different content areas, and several university-based ADLs, Georgia and Pittsburgh, were already a year or more into the development and pilot testing of the assessments for Early Adolescence/Generalist (EA/G) and Early Adolescence/English Language Arts (EA/ELA) certificates, respectively. While the Georgia and Pittsburgh ADLs had already made significant progress in their own assessment development efforts, the NBPTS's Technical Advisory Group (TAG) and the NBPTS itself were still cautiously encouraging experimentation, putting few if any constraints on how the "new" ADLs should go about finding solutions to the inherent challenges of such a complex assessment system.

WestEd's Science ADL received passing words of advice and caution, such as, "Don't make portfolio entries too large or cumbersome," "We can't afford to take half a day to score a single entry," "Videotapes are very telling pieces of evidence and should be central to the portfolio," and "Interviews don't work as assessment center exercises." At this time, there seemed to be many more questions than solutions. In general, there was a tone of "Do your best, share what you know with other ADLs, but keep in mind that this is a high-profile, nationally important project."

In this climate, we began simultaneously developing the portfolio entries and assessment center exercises. The portfolio entries would focus on a teacher's practice in the classroom, while the assessment center would focus on teacher's content knowledge. We had a fairly well-defined structure for developing the portfolio aspect of the assessment, yet parameters for developing the assessment center exercises were less clearly defined.

Based on our literature review and dozens of conversations with science teachers, we created a conceptual framework to guide our work, which addressed important questions such as: How can we be sure to address what is at the heart of science teaching? How can we design the assessment to allow teachers to demonstrate their accomplishments in different, yet equally valid ways? What are the best ways to gather second-hand evidence about a teacher's practice? And most importantly: How can we ensure balanced and thorough coverage of the AYA/Science Standards?

In developing the assessment center exercises we struggled with the fact that the NBPTS's AYA/Science Standards, still only in draft form, did not clearly specify the content knowledge domain, focusing more on instruction and classroom environment instead. We were required to address content knowledge across four scientific disciplines (i.e., biology, chemistry, Earth and space science, and physics), but the depth and breadth were not specified. What the AYA/Science Standards did make clear was that hands-on investigations, problem solving, and inquiry are at the heart of teaching all science, regardless of the specific discipline.

Given these guidelines, we needed to determine the implications for the assessment center exercises. In particular, we considered whether it was feasible to have teachers manipulate tools and equipment at an assessment center. Could a computer simulation satisfactorily replicate the kind of thinking that occurs in a laboratory setting? In these early months, we explored several different computer-based simulations, developed prototype exercises, and made substantial progress in defining the domain of content knowledge required.

Then, within eight months of our having begun development, the NBPTS halted development efforts across all the ADLs because its resource development efforts had not kept up with the costs. We no longer had funding to continue our work.

While our development process was interrupted, it was not completely halted, and in general we followed the steps outlined in Figure 2. This figure provides an overview of the entire assessment development process and participants.[2]

<Insert Figure 2 about here>

Assessment Development: Take Two

Eager to resume our work, in 1995 we co-authored with the NBPTS a proposal to the National Science Foundation (NSF) and successfully obtained funding to continue development. In the interim period, the NBPTS had restructured its entire development process, opting to replace the ADLs with one general contractor to develop, revise, and administer the assessments for all certificates. By the time we received our NSF funding, we found ourselves the only remaining ADL, with the Educational Testing Service (ETS) now responsible for all the other certificates.

Thus, this second stage of development brought with it a changed context. The NBPTS now placed greater emphasis on the timely development of new assessments based on a standard structure and did not encourage experimentation. An emphasis on standardization emerged as a way to ensure efficiency in both development and administration, theoretically allowing the NBPTS to develop and offer a broad range of certificates more quickly. The NBPTS believed its new standardized structure for assessment development struck a balance between addressing technical concerns and keeping administration costs reasonable (e.g., reducing scoring time).

By late 1995, the AYA/Science Standards had also been finalized, and the TAG had been phased out. In addition, the NBPTS had found that significantly fewer minority candidates were becoming NBPTS certified in those certificate areas already offered. This "adverse impact" became a development concern, and we had the additional charge of ensuring that the assessment neither discouraged nor unfairly disadvantaged any group of candidates. To complicate the situation further, a nine-month delay in actually receiving the NSF funding,[3] combined with pressures to make the AYA/Science certificate available as soon as possible, necessitated greatly compressing the remaining development cycle. Our original development timeline of 18 months for the portfolio was now reduced to 12 months. At about the same time, with great sadness, we experienced the death of Mary Budd Rowe, whose vision and style had guided much of our work.

As we moved ahead in this changed context, we struggled most with identifying those areas of the proposed assessment system that were now defined by the NBPTS and those areas that were still open to our innovation. Recognizing that the NBPTS had now made some decisions about the general development and standardization of the certification system, we were uncertain as to which elements, if any, we could continue to shape. For example, regarding the portfolio, we were told by the NBPTS to follow a template with five portfolio entries, including two entries based on student work and written commentary as evidence, two other entries based on video segments and written commentary as evidence and one entry focusing on teachers' documented accomplishments outside of the classroom. However, we wondered how much latitude, if any, remained for including different response formats in the portfolio. Also, were the format and sequence of the entry directions fixed? Could lessons learned from the pilot that might deviate from the standard template be incorporated into the final portfolio?

We all had gone into this project hoping that we could develop an innovative, state-of-the-art assessment that then could be standardized in its implementation. Many of us had been attracted to this project because of the opportunity to innovate. This sudden shift in the NBPTS's expectations seemed to fall back on old paradigms. Could we resign ourselves to maintaining these old paradigms?

With respect to the assessment center exercises, many of the NBPTS's new development parameters also differed from our earlier work in 1994. For example, the NBPTS had determined that all assessment center tasks should be completed over six hours in one day, something we had originally proposed and that was later supported by the NBPTS's experience. Another change was that the assessment center exercises would be delivered via computer, on site at a Sylvan Technology Center (of which there are more than 200 across the United States). As with the portfolio development, we found ourselves wondering if the drive for standardization left any opportunity for innovation and, if so, where, and to what degree. While the NBPTS was moving toward having the assessment center element of each certificate be composed of four 90-minute, written essay exercise blocks, we were hoping that one of these blocks could contain computer simulations to serve as a substitute for hands-on investigations. Would there be any place for lab simulations? How would teachers respond to the heavy emphasis on writing? How would teachers respond to having to take the entire assessment on a computer given the likelihood of their having different levels of computer experience?

This, then, was the shifting context in which we carried out the work of developing the AYA/Science assessment. Each of the following three sections contains several illustrations of some specific lessons learned in developing the portfolio entries, the assessment center exercises and the scoring system, respectively.

STORIES OF DEVELOPMENT: THE PORTFOLIO

The five stories in this section describe: (1) how we strove to obtain balanced and thorough coverage of the AYA/Science Standards as the portfolio entries collectively changed over time; (2) how the use of a development "shell" contributed to consistency, efficiency, and user-friendliness; (3) how the use of alternative response formats (e.g., audiotape) solved some problems and created others; (4) how the need for creating fair assessments may result in greater prescription and less teacher choice; and (5) how best to assure face validity with science teachers in the field, while addressing the technical considerations of assessment.

Portfolio Creation and Evolution

As we look back over the development of the AYA/Science portfolio, we jokingly call it the “incredible expanding-shrinking portfolio.” That is, the portfolio’s size has changed substantially over time in response to changes in current wisdom, a changing political climate, competing demands of the APPLE[4] criteria, and, ultimately, which stakeholder(s) had final say. Despite this tug-of-war, our bottom line was to ensure balanced and thorough coverage of the 13 separate AYA/Science Standards.

In the early months of the first year of development (Spring 1994), we reviewed the work done by others, convened conversations with dozens of leading educational researchers, and consulted research findings to create an internal document or “conceptual framework” that provided theoretical guidance for the portfolio development. This conceptual framework provided working definitions of our tasks, spelled out different approaches used in assessment development, explored research findings about the differences between expert and novice teachers, and helped us to clarify the skills and types of teacher knowledge we wanted to target in the assessment.

On a theoretical level, we thought about the development of the portfolio in terms of three components: one or more *tasks* that require teachers to demonstrate what they know and can do in teaching science; the *response format* in which science teachers provide evidence of their competencies; and the *scoring system* by which candidates’ performances are judged (Ruiz-Primo & Shavelson, 1996; Solano-Flores & Shavelson, 1997). We intended the portfolio to provide a rich picture of a teacher’s abilities in the context of his or her own classroom, and here we took guidance from Shulman (1998, p.34), who described a school-site portfolio as a “structured history of a carefully selected set of coached or mentored accomplishments substantiated by samples of student work and fully realized only through reflective writing, deliberation, and serious conversation.”

One of the first most important things we did was to bring together a team of science teachers, including teachers from the AYA/Science Standards Committee and the WestEd development team to breathe life into the science portfolio. We began by trying to answer questions such as: What is essential to good science teaching? What is unique to teaching science? How can the AYA/Science Standards guide our choices?

Weeks later, when our list of ideas, shared experiences and visions covered chart paper on several walls, we divided into smaller groups and worked on consolidating these ideas into the titles and content of actual portfolio entries. After much conversation and some compromise, the groups came to an agreement on the essence of five portfolio entries. Specifically, at that time, we determined that the entries would focus on the following: (1) teaching major ideas in science; (2) assessing student work; (3) hands-on scientific inquiry; (4) professional commitment; and (5) portfolio reflections and science teaching philosophy (Figure 3). The 13 AYA/Science Standards were distributed across these five entries in a way that to some extent allowed for multiple coverage of each standard. This provided teachers with the opportunity to illustrate how they met the standards in more than one way and more than one teaching situation.

<Insert Figure 3 about here>

Several months later, things changed—yet again. In October 1994, the half-dozen ADLs were invited to Washington, DC to share lessons learned, discuss development challenges, and hear TAG members present their findings from earlier development work. At this meeting, the ADLs were asked to increase the number of entries in each portfolio to between nine and 12, with the expectation that doing so would increase the dependability of the assessment scores. From a practical perspective, this was encouraging because the scoring time for each entry would be greatly reduced, and in theory, the overall scoring time would also be reduced.

For our ADL, this was nearly double the number of entries we had been developing. Along with the other ADLs, we went back to the drawing board with our teacher developers, asking ourselves, which of our existing entries were relatively large? Were there natural divisions in these existing entries? Did we need to develop new entries to fill holes in our sampling of the AYA/Science Standards? The answers to these questions led us to develop an expanded set of nine portfolio entries (see the second column of Figure 3). And, in fact, our teachers, by and large, felt that these slimmer and more tightly focused entries were more manageable to assemble.

Yet one year later (1995), after the NBPTS had hired ETS to replace the other ADLs, a still newer set of portfolio development guidelines was adopted. In the intervening year, the costs associated with developing, delivering and scoring its assessment package had emerged as a major concern for the Board. A cost analysis had indicated that the time involved in scoring assessments was the major expenditure. To keep costs down and to reduce the workload for teachers, the NBPTS now recommended that the number of portfolio entries be reduced to six. In addition, both the Board and ETS wanted standardization across certificates. Moreover, the new structure prescribed two student work-based entries and two videotape-based entries, all of which focused on the particular content of the certificate, and two additional entries that were identical for candidates irrespective of their content specialty. Our challenge was now to pare down to six the nine portfolio entries we had already developed while simultaneously ensuring coverage of the AYA/Science Standards.

To do this, we rebuilt the matrices to look at the content coverage and distribution of the AYA/Science Standards across the different entries. It was easy to fall into the trap of looking for every standard in every piece of evidence a teacher provided; however, to better guide teachers in selecting their most compelling evidence, we needed to focus each entry narrowly around a manageable number of standards. But what constitutes a manageable number of standards to be addressed in each entry? What evidence are teachers being asked to collect in each entry? What type of artifacts provides the best evidence for each standard?

Each reconceptualization of the portfolio required that we carefully map back the entries onto the AYA/Science Standards to ensure internal consistency among the focus of the entry, the description of the nature of the entry, the standards being addressed, the questions teachers would respond to in the written commentary, the explanation and advice with respect to how responses would be scored, and the scoring system (e.g., rubric). Figure 3 provides a comparison of the portfolio entries and AYA/Science

Standards coverage as the entries evolved throughout development. Figure 4 provides a summary of the final AYA/Science portfolio entries.

<Insert Figure 4 about here>

We had begun developing the portfolio with the idealistic notion that it would be an assessment written by science teachers for science teachers that aimed to cover the standards. However, coverage of the standards proved to be only a small, albeit important, consideration in the development process. In reality, during quarterly review sessions, the development of the portfolio was influenced by select members of the NBPTS staff, a couple of staff members representing the general contractor (ETS), and several NBPTS Board members—each bringing his or her own perspective about criteria and priorities. Much later, our work was also reviewed by the AYA/Science Standards Committee and approved by the NBPTS Board of Directors. We inevitably had to balance multiple needs, multiple opinions, and often competing components of the APPLE criteria. To name but a few, these concerns included teachers' needs, the need for standardization to lower scoring costs, and making the assessment equitable. To balance these various influences and to be most efficient in a climate of continuous change, we found that we had to be flexible, patient, and resilient. This meant being willing to go back to the drawing board repeatedly, focusing on the exciting intellectual challenges rather than the mundane frustrations.

"Shells" for Standardization and Simplification

In developing the AYA/Science assessment, we were aware of the importance—and challenge—of coming up with cost-efficient procedures for developing the portfolio entries. We intended to honor the "E" (economic affordability) of the APPLE criteria. Being cost-efficient meant finding ways to simplify and standardize the development process to create consistent portfolio entries.

To meet this challenge, we drew on experience we had gained on other projects using "shells" to develop multiple constructed-response assessments (Solano-Flores & Shavelson, 1997). In the context of alternative assessment, shells serve as blueprints, or templates, for efficiently generating assessments in a short period of time by providing step-by-step guidelines for assessment developers (Solano-Flores, Jovanovic, Shavelson, & Bachman, 1999). In this particular context, a shell serves as a development tool that outlines the separate sections of each portfolio entry and describes what type of information should be provided, where that information should be provided, and in what format. This would allow for consistency in the content, sequence, and format of information across all portfolio entries (see Figure 5 for an example of a shell that served as an evolving document throughout development of the AYA/Science Portfolio).

<Insert Figure 5 about here>

The use of a shell in the AYA/Science assessment simplified the development of the complex portfolio entries that involve a variety of products (e.g., narratives, videotape footage, samples of student work). Moreover, to ensure standardization across a wide range of teacher candidates with different backgrounds who teach in very different contexts, and to eliminate adverse impact, the directions to the teacher candidates must be long, precise and exhaustive. For every entry, these directions must specify what actions the teacher candidates must take to submit their responses and the format criteria that the candidates must adhere to (e.g., page limits, font size, specifications for videotape footage). The directions must even anticipate possible misunderstandings from the candidate and provide directions on what not to do (e.g., "Do not submit more than five pages; assessors will not read more than the text length specified").

The shell we developed consisted of a set of directions specifying the sections that every entry should contain and the sequence in which those sections should appear. As with other forms of alternative assessment, the portfolio entries underwent pilot testing through an iterative process of try-out, review, and revision (Solano-Flores & Shavelson, 1997). In this process, the content and format of the shell also evolved and became more detailed, to the extent that it even specified where page breaks should be inserted.

This approach of updating the shell, based on our experience of trying out and revising the entries, had two unanticipated benefits. First, the shell became a document that formalized our thinking at each stage of development about the portfolio and the nature of knowledge and skills we intended to address. Second, in revising the entries for every new pilot test, we frequently found ourselves using the shell as a frame of reference; it became a succinct description that summarized how we wanted to format and sequence the portfolio. Thus, although the shell could not tell us much about the content of the entries, it certainly helped to make our discussions more systematic. This allowed us to think carefully about the content of the portfolio without having to reconceptualize how to present that information over and over each time we tackled a new entry.

Due to time constraints, we wanted to develop the portfolio entries simultaneously, assigning each entry to a different team member for development and revisions. But we knew the potential problems. In reviewing other portfolios over time, we had noticed that entries within a particular portfolio might vary considerably in format and even style, which could, in turn, make it difficult for a teacher candidate to understand how one entry differed from another in terms of content and focus. Using a shell ensured that we could develop entries simultaneously *and* have continuity across all entries, eliminating any potential confusion for teachers once the assessment became operational. A common structure across entries, then, not only facilitated the process of portfolio development but also ensured assessment fairness. Candidates could navigate through the separate portfolio entries more easily when they only had to become familiar with one user-friendly format and one sequence of instructions.

The use of a shell thus became a critical tool for the development process—for both the portfolio and the assessment center activities as that work progressed. Its role was especially important given the magnitude of the project, the great number of developers, and the changing expectations.

Exploring Alternative Response Formats

Developing the first portfolio entry (Entry 1: Teaching Major Ideas Over Time) was especially interesting because it raised a number of complex development questions, which in turn provided us with important lessons. This was due in part to the nature of Entry 1 and in part to the fact that it was the first to be fully developed. To understand some of these questions and challenges, we provide next an overview of Entry 1, and we then describe the development considerations we faced.

Figure 6 is an excerpt from Entry 1 that describes the nature of the entry and provides teachers with a summary of what they must submit. We developed this entry to reflect the important notion of teaching a major conceptual unit in science and to provide a holistic look at teaching by focusing on an extended instructional period (i.e., a minimum of three weeks). Entry 1 was designed to complement the videotape-based entries that show more discrete pieces (e.g., 20-minute teaching moments). Other portfolios had not addressed this “over time” aspect, yet, the art of weaving instructional activities together to develop students’ understanding of science concepts is essential to good science teaching.

<Insert Figure 6 about here>

In addition, our intent was for Entry 1 to capture teachers’ reflective thoughts throughout an extended period of instruction, thereby allowing assessors to gain an understanding of how accomplished teachers monitor and adjust their instruction to better address students' needs. Capturing teachers' thoughts as they teach and not merely their reflections weeks later was an important objective, but not easy to accomplish. The AYA/Science Standards place high value on teacher reflection; Standard XIII: Reflection states, "Accomplished science teachers constantly analyze, evaluate and strengthen their practice in order to improve the quality of their students' learning experiences." Thus, each portfolio entry requires teachers to reflect on their practice. Moreover, current educational research indicates that the degree to which instructional decisions are based on deliberate rationale or intuition is a continuum, with accomplished teachers generally on the higher end and novice teachers on the lower end (see, for example, Berliner & Gage, 1989).

Teachers reflect informally about their teaching all of the time. Yet given the great importance of reflection, we felt we needed to formalize the reflective processes. Thus, each entry contains questions that probe teachers' reflective practice. In our first attempt to capture teachers' reflective thoughts during the period of instruction, we asked them to keep a daily log of class activities along with their thoughts about successes, failures and next steps. From this log, they would select several examples to submit with their portfolios. Although this idea was conceived by teachers from the development team, other practicing teachers who reviewed the entries in focus groups were generally not receptive to the idea. They made comments such as, "I do this kind of reflection all of the time, but I never write it down. When would I find the time during the day?," "I'm sure some people like keeping journals, but I'm not one of them," and "I'm a science teacher, not an English teacher."

We were getting a clear message that, although a written log might be a solution for teachers of English or Language Arts, many science teachers were resistant. In fact, as science teachers looked at the portfolio as a whole, they were overwhelmed by the writing demands. Interestingly, other teams who were developing assessments in different content areas found similar concerns. Although most teachers and development teams agreed that accomplished teachers should also be competent writers, they were concerned that the assessment might more directly measure their written communication skills rather than their skill in teaching. In particular, science teachers questioned whether they would be able to adequately represent in writing what they were able to accomplish in the classroom.

These concerns pushed our team to ask: What alternative response formats would serve the same assessment purpose and allow teachers to utilize different modes of communication? We considered a novel substitute—allowing teachers to use an audiotape to record their reflections. The assessment could provide guiding questions that teachers responded to via tape recorder. These questions would capture their expectations prior to the period of instruction, along with their mid-course thoughts and end-of-unit reflections. But, while lessening the writing demand on teachers, the use of audiotape also raised as many questions as it addressed. Would classroom teachers like it? Would talking into a tape recorder be foreign or uncomfortable? What technological challenges and additional burden would audiotapes present? What were the scoring implications? Would it be necessary to transcribe candidates' responses? Would scoring be more time consuming and more costly, and outweigh the benefits gained? Would alternative biases be introduced? Could the use of an audiotape offer a means to address issues of adverse impact documented among specific teacher groups? Would honoring teaching as an oral profession lend greater professional acceptability? Would the use of an audiotape support teachers' own assessment efforts to address their students' multiple intelligences? Would audiotaping actually reduce the amount of writing or would teachers write their responses and then read what they wrote into the tape recorder?

Audiotaping clearly introduced unknown variables, but it also promised solutions to some important issues. Could videotape—a better known entity because it was already required in other portfolio entries—be used to record teachers' reflections? With videotape, teachers could both "show" and "tell" their responses. What were the assessment implications? How is using video to capture teachers' reflective thoughts different from taping a class while the teacher is teaching? Although little research existed around the use of videotaping or audiotaping for this purpose, we were interested in exploring the possibility.

In agreement with the NBPTS and the general contractor, we pilot tested Entry 1 in all three response formats—a subgroup of teachers would respond to questions in writing, another group would audiotape their reflections, and a third group would respond on videotape. Although this pilot test did not have the design of a formal study, it yielded exploratory information as to the feasibility of these modes, both with classroom teachers and with the assessors during a pilot scoring session.

Careful review of a small pilot test sample of teachers' portfolio entries, along with their responses to questionnaires and interviews, showed that teachers were willing and able to record their reflections using tape recorders and video cameras. Several of the teachers reported being somewhat uncomfortable talking into a tape recorder, while others liked the idea. However, talking "to" a video camera seemed artificial and left some teachers more uncomfortable than being videotaped while teaching their classes. Although teachers were prompted to provide "show-and-tell" videotaped responses (e.g., by showing three-dimensional student products or visual teaching props), we found that all of the teachers who used videotape gave a "talking-heads" response (e.g., each teacher sat and talked squarely into the camera), thereby providing little or no additional information from that given on an audiotape.

There were many similarities between scoring an audiotape and a videotape. Due to scoring time constraints and logistics, assessors would not be able to rewind a tape and view or listen to any section more than once. Assessors would need to make use of a note-taking sheet to keep careful record of evidence presented in the tapes. While audiotapes would allow assessors to refer to other pieces of evidence (e.g., student work samples) while listening to the teachers' explanations, videotapes required continuous viewing. Many questions were raised about the use of audiotapes and assessor bias. However, our experience in training assessors to view a videotape and make a fair evaluation gave us confidence that it would also be possible to address issues of bias (e.g., regional accents) introduced by an audiotape (see the section entitled, "The Power of Videotape Footage in Bias Training").

Again, because this was merely an exploratory, not a formal, comparative study, we could not conclusively answer many of our initial questions. For example, we had no information as to whether or not audio- or videotape might address the concerns of adverse impact or significantly reduce the writing demands for teachers. Although scoring times were comparable between the audiotape versions and the written responses, we could not guarantee that it would not take longer to score the audiotapes, especially if producing a written transcript were deemed necessary.

These unknowns, along with possible cost increases, pressure to have operable assessments in the shortest possible time, and the concerns of working under a compressed development timeline tilted the scales. The NBPTS chose to use a well-established response format (i.e., written responses) until someone could conduct further research on alternative response formats (e.g., audiotapes). Although we could not disagree with this decision, we regretted that the short timeline and economic restrictions prevented us from further investigating the possibilities and limitations of this form of assessing teachers. In this instance and others, we were continually reminded that innovation requires sufficient time and resources.

As we also discovered in developing this entry, that the development process often involves a series of trade-offs. As with any task, portfolios will favor some teachers over others depending on their ability to articulate in writing versus other response formats (whether graphic or oral, for example). Both in our interaction with the NBPTS and within our team, there was always a tension between creativity and methodological soundness. Certain solutions may create other problems, and often there is no one right answer.

The Tension Between Individual Creativity and Measurement Necessity

Our development of the AYA/Science assessment confirmed that measurement concerns sometimes outweigh allowances for individual difference and creativity. We experienced a constant tension between wanting to develop open-ended portfolio entries that allowed for a greater latitude of responses and the need to restrict both the content and the format of teachers' responses with highly specific and prescriptive prompts.

Creative, inspired, and accomplished science teachers are likely to come up with very different responses to an open-ended assessment. In high stakes instances such as NBPTS certification, this is not always a good thing. An assessment must balance competing objectives such as soliciting concrete evidence that is readily comparable across candidates while at the same time allowing appropriate leeway for individuals to respond in a manner that accurately portrays their best teaching. Below we describe the evolution of the portfolio entry directions for the "Instructional Context" section of the written commentary as a small example of this tension between open-endedness and specificity.

All portfolio entries require a written commentary. It is perhaps *the* key opportunity for candidates to supply evidence that they are meeting the standards addressed by a particular entry. This commentary is where the candidate can explain connections among disparate pieces of evidence (e.g., student work samples or videotape footage), communicate his or her insights, and share pertinent information found nowhere else in the NBPTS assessment (e.g., student background data).

A crucial part of the written commentary is the segment in which candidates are asked to supply background information that will help assessors understand arguments and justifications made by the candidate later on in the commentary. This segment, which is called the "Instructional Context," has an initial prompt that is fairly uniform across all entries. An examination of this prompt in early and late versions of the portfolio entries illustrates how we had to "close the box" by prescribing the information that candidates must include in their Instructional Context response so that the performance of all candidates would be assessed on an equal basis (see Figure 7).

<Insert Figure 7 about here>

The early version of the Instructional Context prompt is suggestive and inclusive. Phrases such as "This might include," and "You might wish to add" lend, superficially at least, a degree of open-endedness to what the candidate is being asked to do. One could argue that, in a high stakes environment, any "suggestion" should rightly be interpreted as a "directive."

In pilot tests of early versions, some teachers appeared to take the "hidden directive" approach. They supplied certain "might wish to" information that others did not and subsequently fared better in scoring than those who interpreted the prompt more literally. Accordingly, we attempted to level the playing field by revising later versions of the prompt to specifically request that *all* candidates include certain information such as title, subject matter, and size of the class. These particular details, while seemingly obvious and mundane, were important in understanding the student work or teacher's choices; nevertheless they were absent or not readily discernible in some teachers' responses.

We retained a certain degree of open-endedness in the prompt by wording the final portion to read, "and any other realities of the social and physical teaching context … that are relevant to this entry." Although some guidance is given in the form of possible "realities," (e.g., available resources, ethnic and linguistic diversity), candidates are free to choose those that fit their own particular situation. We, as assessors, cannot know all the contextual details that should be included. Because of this, the box must remain "open" and candidates must carefully choose what information to include.

As this example shows, the "open" versus "closed" box poses an unavoidable and significant dilemma for developers. When designing any assessment, developers must consider the limits of candidate choice. We chose to temper our decisions in this regard by using a measurement-oriented fairness rule: if candidates' responses are likely to be scored more accurately if they are specifically requested to do something, then ask them to do it. Some may recognize this rule as a variant of the "TTWYW" or "Tell Them What You Want" admonition. The importance of this axiom should not be underestimated, especially in instances of high stakes assessment.

This was an important lesson, all the more so because we found it counterintuitive. We had thought that allowing teachers to respond more openly to a prompt would allow for greater fairness in their expressions of accomplishment and, ultimately, in scoring.

However, this openness actually led to greater interpretation on the part of the candidates and, thereby, induced greater inequity.

Involvement of Science Teachers in Assessment Development

We knew that the substantial involvement of practicing science teachers was key to successful assessment development. Yet, we wondered how best to involve teachers in a substantial, rather than superficial, way. We did not want to squander classroom teachers' efforts on the time-consuming stages of writing, "wordsmithing," and polishing the assessment, or on addressing issues of consistency, adherence to policy, and technical assessment considerations.

Practicing teachers were particularly valuable given their keen knowledge of the audience, their advice for making the assessment relevant, and their subsequent advocacy on behalf of a project they believed in. For example, Mary Budd Rowe was concerned that the assessment should immediately excite and engage science teachers. In developing the assessment, we imagined what it would be like to be a science teacher receiving this assessment package in the mail. What would a teacher read first? How would each portfolio entry be introduced?

It was easy for our team of teacher developers to understand the experience of teacher candidates, and, consequently, they had dozens of ideas for how best to begin each entry so as to be inviting to teachers. We experimented with inspirational quotes and teacher-written, personal vignettes, but in the end the NBPTS's standardized format prescribed a digested summary of what is expected of accomplished teachers (see Figure 8).

<Insert Figure 8 about here>

We attempted to make choices and employ a process that balanced science teachers' values, voices, and needs with the technical considerations of the assessment. We found that teacher developers on our team greatly assisted us in assuring the appeal and face validity of the portfolio to classroom teachers. We found that the best way to maintain an allegiance to teachers was to ensure that they were central to the process of development. This means providing the time and resources to ensure that teachers are present and heard at all stages of development.

This does not imply that even the most accomplished classroom teacher has the experience, skills or time to be involved in every aspect of development. However, on our development team, teachers were invaluable in generating ideas, reviewing materials, pilot testing the assessment, and participating in scoring. Teachers on the development team were also excellent liaisons between assessment experts and other teachers. They became excellent ambassadors for the NBPTS, and contributed to the professional acceptability of the assessment. Development team teachers facilitated focus group discussions and, as teachers themselves, could readily hear what would resonate with other teachers. It was valuable that our core leadership team was composed almost entirely of past and present science teachers who worked alongside a diverse group of top-notch practicing teachers.

STORIES OF DEVELOPMENT: THE ASSESSMENT CENTER

The two stories in this section highlight issues relating to the development of the assessment center exercises. The first describes how we defined the content knowledge that an accomplished high school science teacher should possess. The second describes how we designed the assessment center activities to capture a teacher's depth and breadth of knowledge within a vast content domain. It is important in reading the following section to remember that the NBPTS had determined that the total time for administering the assessment center exercises should not exceed six hours.

Defining Content Knowledge

In developing the assessment center component of the science assessment, we faced interesting challenges in defining the vast and complex knowledge domain—the "Knowledge of Science." As noted before, although the AYA/Science Standards placed importance on teachers' subject matter knowledge, they did not specifically define the content knowledge science teachers should possess (see Figure 9 for excerpts from Standard II: Knowledge of Science). Further complicating this issue, the science certificate issued by the NBPTS is the same for all science teachers, regardless of their specialty area (biology, chemistry, Earth and space science, physics).

<Insert Figure 9 about here>

Exploring how best to define the domain of science content caused us to question what grain size of knowledge was appropriate—at what level of depth and breadth should teachers understand content within each of the four specialty areas?

To address the first challenge in developing the model for the assessment center—the limited specificity about content knowledge in the AYA/Science Standards—we turned to nationally recognized documents that described content standards for *students*, such as the *National Science Education Standards* (National Research Council, 1996), *Benchmarks for Science Literacy* (American Association for the Advancement of Science, 1993), and the *Science Framework for California Public Schools K-12* (California Department of Education, 1990). At the time of our work, no documents existed that outlined content standards for science *teachers* (other than the AYA/Science Standards). Therefore, we created a document that combined the AYA/Science Standards and the other national standards documents that served as a framework for outlining the content and organization of assessment center exercises.

To address the second challenge, the complexity of having multiple disciplines within the science knowledge domain, we synthesized the various national and local standards documents for grades 9-12 and organized their content into the areas that are typically taught in the United States—biology, chemistry, Earth and space science, and physics. The result of this synthesis was a set of five domains (the four mentioned previously and an additional, cross-disciplinary domain referred to as Common Aspects of Science) subdivided into core concepts (see Figure 10 for an outline of these core concepts; these core concepts include a number of subtopics, which we have not included here).

<Insert Figure 10 about here>

While perhaps obvious, it is worth highlighting the importance of knowing what you want to assess before trying to develop exercises that measure achievement of that knowledge or skill. In determining what we needed to assess, we found that we had to be resourceful by tapping into other standards documents and synthesizing those documents to delineate the domain of knowledge a candidate should possess.

Assessment Center Design

Our next task was to design exercises that would allow us to assess the scope and level of content knowledge that a candidate possesses. With science, in particular, we faced the challenge of designing an assessment that authentically captured the skills and knowledge central to science—namely, scientific inquiry in the lab. The assessment center was designed to assess a candidate's *depth* of knowledge with exercises that addressed an area of emphasis that a candidate could choose (e.g., biology, chemistry, Earth and space science, or physics) along with Common Aspects of Science[5] (e.g., the methodological, philosophical, and social dimensions critical to all scientific disciplines). We assumed a minimum of first-year, college-level knowledge for all the exercises we developed.

In addition, a candidate's *breadth* of knowledge was to be assessed by exercises that focused on the three complementary comprehensive areas outside the candidate's chosen area of emphasis. Thus, for example, a candidate whose chosen emphasis is physics would be assessed in depth in physics and Common Aspects of Science, whereas the breadth of their knowledge is assessed in the remaining three comprehensive areas (biology, chemistry, and Earth and space science; see Figure 11).

<Insert Figure 11 about here>

In collaboration with NBPTS and ETS, we also introduced the notion of "depth-breadth of knowledge" to address the knowledge that is related to establishing connections across content areas around basic (universal) ideas in science. More specifically, depth-breadth of knowledge refers to a teacher's capability to understand how a basic idea such as the relationship between form and function—whose principles are the same regardless of content area—applies to the candidate's area of emphasis and how that idea can be identified in phenomena that belong to complementary content areas.

In addition to addressing the design challenges related to defining content domain, we recognized that different methods for measuring knowledge tap into different aspects of academic achievement (Baxter & Shavelson, 1994; Dalton, Morocco, Tivnan, &

Rawson, 1994; Ruiz-Primo & Shavelson, 1996). Therefore, to achieve dependable scores, following the suggestion made by Resnick and Resnick (1992), among others, we combined different types of exercises, from short answer exercises to constructed-response exercises—such as essay items or performance tasks.

Given all this, we formalized an assessment center structure that was consistent with the four 90-minute blocks prescribed by NBPTS (see Figure 12). A candidate's depth of knowledge was assessed in two 90-minute blocks of exercises that addressed both content and pedagogical content knowledge, including essay exercises intended to elicit inquiry-and-analysis skills, student work analyses, and laboratory investigations intended to elicit science process skills. The breadth of knowledge exercise block consisted of brief essays. Exercises that addressed depth-breadth of knowledge consisted of three 30-minute essay exercises that required the teacher to establish connections across content areas using a given basic idea in science.

<Insert Figure 12 about here>

Common Aspects of Science were assessed in the context of a teacher's area of emphasis. For example, a physics teacher might be asked to discuss the implications of the invention of the laser with respect to science, technology and society. This ensures that reasoning about societal issues of science (technology and environment) and investigation skills are assessed within a context that is most relevant to the teacher.

One of the bigger challenges we faced was how best to assess science process skills in the absence of authentic, hands-on laboratory investigations. Originally, we designed investigations that were carried out with computer simulations of processes in science that candidates could manipulate. We believed that, although computer simulations are not perfect surrogates for authentic hands-on tasks (nor are paper-and-pencil investigations), the administration of computer simulations would be less expensive and logistically simpler than hands-on tasks. In addition, computer simulations could bring more validity to the tasks included in the assessment and elicit a kind of knowledge highly valued by science teachers.

We used Science Explorer 3.0, a personal computer software package developed and published by LOGAL Software, Inc. to develop the original investigations. This software provided a series of interactive, inquiry-oriented explorations in science. With these simulations, we developed performance tasks in which candidates investigated phenomena by manipulating variables and observing the results of their actions. Due to their innovative nature, we attempted to ensure that these simulations would not privilege candidates who had experience with computers or certain problem-solving styles.

Despite, and in part *due to,* the innovative nature of these exercises, the NBPTS decided not to include these simulations in the assessment. That determination may have been driven in part by the decision to deliver the assessment center exercises through the Sylvan Technology Center computer operating system because it could have been difficult to align the architecture of the simulations with ETS' existing proprietary operating system. We did, however, transform these simulations into paper-and-pencil prompts, building on the knowledge we gained in developing the actual, “hands-on” simulations.

In the end, we were reminded of the lesson learned when developing portfolio entries—innovation requires a significant commitment of time and resources, and in the period of innovative uncertainty where new questions emerge, difficult decisions must be made. While the use of cutting-edge technology may have provided a solution for assessing depth of teacher’s knowledge and science process skills, technology also amplified uncertainty, risk and resources needed for assessment development.

STORIES OF DEVELOPMENT: SCORING

Our charge in developing the AYA/Science assessment included developing a scoring system. Because of the documented “adverse impact” that appeared to affect certain groups of candidates, we believed that the scoring system should include a bias training component to help reduce the impact of potential assessor bias. The story in this section highlights how powerful bias training can be with the aid of videotape footage and dyad discussions around that footage.

The Power of Bias Training for Scorers

Bring a group of teachers together from all over the country and everyone will have a different opinion about how best to teach a class of students. Teachers hold passionately to their ideas and argue vehemently about the merits of how and what should be taught. During scoring sessions that we facilitated to assess the usability and appropriateness of the scoring rubrics, we encouraged the 15 or so participating teachers to draw on their professional knowledge to make informed scoring decisions, while simultaneously recognizing bias that emerges as a result of differences in each of our teaching experiences and contexts.

To address the potential problem of bias, we brought this issue to the scorers' attention, talked about the ways in which strong emotional reactions can be an indicator of bias, and emphasized the need to consider only the rubric criteria (rules for scoring) when making scoring decisions. We also used diverse performances for training scorers, hoping that, through discussions about these performances, we would flush out many of the biases that could result in an adverse impact on any particular group of teachers.

Despite these efforts, we witnessed multiple instances of bias on the part of mainstream scorers when looking at performances by non-mainstream teachers in non-mainstream school environments. For example, one scorer said, "You can't expect much from kids in that setting anyway," as many other scorers nodded in agreement. Obviously, such lowered expectations of students results in a distorted view of many successful classrooms. Similarly, in response to a videotape of an African-American teacher from New York City, many scorers described his performance as "ranting and raving." Others defended his communicative style, saying that he was "preaching" to his students in a way that is appreciated and understood within African-American culture.

This videotape seemed like a clear example of "culturally informed practice" that was not viewed equitably by mainstream scorers.

Overall, in these early pilots of national teacher assessments, we witnessed many instances of miscommunication among scorers as teachers tried to interpret what was happening in school environments different from their own. Even though the leaders at these scoring sessions were themselves a diverse group, they did not have the tools to educate people about these differences or to help the teacher scorers become more cognizant of their own culturally informed point of view (Ladson-Billings, 1995) and the bias that results from assuming that one's own communicative style is, as Lisa Delpit says in her book *Other People's Children* (Delpit, 1995), "simply 'the way it is.'"

Through these early experiences, we came to several realizations: (1) telling mainstream scorers to be aware of bias was not enough; (2) mainstream teachers were in many cases unfamiliar with non-mainstream learning environments and the communicative styles and patterns within these environments; and (3) videotapes of teachers' performances have the power to evoke bias in scorers when they encounter unfamiliar environments. Knowing that bias had to be revealed before it could be transformed, we decided to use videotapes of teachers in diverse settings to evoke bias, discuss reactions, then provide information that could inform scorers so they might more fairly and effectively interpret other teachers' performances.

In our final pilot, we carefully selected short (e.g., 4-6 minute) segments of videotape showing six teacher performances from widely different contexts, both mainstream and non-mainstream. As a rule, each of the selected performances must have been scored highly (i.e., at level 3 or 4 in the 4-point rubric) in an earlier pilot. While viewing each videotape segment, scorers wrote down their reactions to the teacher's and students' actions, the students' demeanor, and the interactions between students and teacher. After viewing the videotape segment, each scorer evaluated his or her response to the performance, discussing his or her feelings about the style of interaction and his or her analysis of the effectiveness of the interactions.

Scorers then paired up to discuss their responses with another scorer, using a strategy called a dyad. In a dyad, one person speaks uninterrupted for a short period of time (e.g., two minutes); then the other person speaks uninterrupted for the same length of time. The dyad rules also require that you listen closely, always maintain confidentiality, never interrupt, and never judge what is being said (Weissglass, 1995). Through this dyad process, at least two things were accomplished: (1) scorers verbalized their own reactions to another person, thereby giving those reactions more conscious reality, and (2) scorers heard a second perspective, often one that differed from their own. This second perspective in many cases surprised people and had the effect of making one aware that one's own interpretation can be highly "uninformed" or is only one interpretation among other equally valid interpretations.

Scorers went through this process six times with the six different videotapes. Afterwards, we met as a large group to discuss reactions to the videotapes and the process. Many noted how comfortable they felt with teachers who mirrored their own style of interaction and how uncomfortable they felt when asked to interpret something that differed in terms of style. At this time, an expert on cultural patterns of communication and the leader of our sensitivity review team, Sharon Nelson-Barber, answered questions and discussed some of the features of these different performances in terms of use of authority, use of language, emphasis in instruction, and other qualities that differ across cultural settings.

Though this videotape sequence was only one of four parts of the bias awareness training, many scorers said it was the most powerful and transformative part for them. In fact, many felt that it opened a new door, and they wanted to see even more videotapes of different teaching contexts and to gain even more information about different patterns of communicative style.

Mainstream teachers may not know what is outside of their experience and may have difficulty interpreting different cultural contexts. Yet, since teachers in general are communicators and observers of interaction, and as most pride themselves on this skill and knowledge, they all found the videotapes to be a highly valued resource. They also found the overall experience to be transformative, not only in their role as scorers but also in their roles within their classrooms and communities.

Thus, we confirmed that bias training in general, and videotape footage in particular, provided powerful experiences for scorers. Although it is time consuming, it will, in large part, contribute toward making an assessment more equitable.

CONCLUSION

Learning from others' can save valuable time and money. As one would imagine, we gained valuable insights from our own development work and that of other ADLs. The stories we have shared highlight some of the issues that arose for us, and, in conclusion, we have presented some of the more global issues here for the benefit of others.

The assessment development process must be mainly concerned with technical aspects such as content coverage, equity, and psychometric issues, yet the process is shaped by many more forces than just methodological and content considerations. It is a complex social process in which developers must consider and balance the multiple and sometimes conflicting interests, needs, and priorities of all stakeholders. At times, trade-offs are essential. One way to smooth this process is to involve all stakeholders from the beginning and throughout development.

In assessment development as in many other efforts, innovation requires substantial time and resources. One must consider these issues up front in order to appropriately address them when budgeting and setting timelines.

Perhaps most fundamentally, we found it valuable to draw on existing research. Our hope is that our experiences will also shed light on the assessment development process and contribute to the knowledge base.

APPENDIX

REFERENCES

American Association for the Advancement of Science. (1993). *Benchmarks for science literacy*. New York: Oxford University Press.

Baxter, G. P., & Shavelson, R. J. (1994). Science performance assessments: Benchmarks and surrogates. *International Journal of Educational Research, 21*, 279-298.

Berliner, D., & Gage, N. L. (1989). Nurturing the critical, practical, and artistic thinking of teachers. *Phi Delta Kappan*, *71*(3), 212-214.

California Department of Education. (1990). *Science framework for California public schools: Kindergarten through grade twelve*. Sacramento: Author.

Dalton, B., Morocco, C. C., Tivnan, T., & Rawson, P. (1994). Effect of format on learning disabled and non-learning disabled students' performance on a hands-on science assessment. *International Journal of Educational Research, 21,* 299-316.

Delpit, L. (1995). *Other people's children: Cultural conflict in the classroom*. New York: W.W. Norton & Co.

Ladson-Billings, G. (1995). Multicultural teacher education: Research, practice, and policy. In J. Banks & C. Banks (Eds.), *Handbook of research on multicultural education* (pp. 747-749). New York: Macmillan.

National Research Council. (1996). *National science education standards*. Washington, DC: National Academy of Sciences.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. G. Gifford & M. C. O'Conner (Eds.), *Changing* *assessments: Alternative views of aptitude, achievement and instruction* (pp. 37-75). Boston: Kluwer Academic Publishers.

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Rhetoric and reality in science performance assessments: An update. *Journal of Research in Science Teaching, 33*(10), 1045-1063.

Shulman, L. (1998). Teacher portfolios: A theoretical activity. In N. Lyons (Ed.), *With portfolio in hand: Validating new teacher professionalism* (pp. 23-38). New York: Teacher College Press.

Smith, M. S., & O'Day, J. (1991). *Putting the pieces together: Systemic school reform. CPRE Policy Briefs*. New Brunswick, NJ: Consortium for Policy Research in Education.

Solano-Flores, G., & Shavelson, R. J. (1997). Development of performance assessments in science: Conceptual, practical and logistical issues. *Educational Measurement: Issues and Practice, 16*(3), 16-25.

Solano-Flores, G., Jovanovic, J., Shavelson, R. J., & Bachman, M. (1999). On the development and evaluation of a shell for generating science performance assessments. *International Journal of Science Education, 21*(3), 293-315.

Weissglass, J. (1995). It will take more than good will: Avoiding lip service and promoting equity in education. *Multicultural Education, Summer*, 17-20.

Figure 1

Tasks Required of ADLs in Developing NBPTS Assessments

- Work with a NBPTS-appointed Standards Committee in the certificate field as it complete[d] its definition of the domain of knowledge, skills, and abilities that reflect highly accomplished teaching in that field; and engage the Standards Committee in the planning, design and review of performance assessment exercises that are consistent with the definition of highly accomplished teaching specified by the Committee.

- Develop exercises and scoring procedures that assess the construct of accomplished teaching defined by the Standards Committee in the certificate field, and integrate the assessment exercises into complete assessment packages.

- Work with other ADLs, the Technical Analysis Group (TAG),* the Field Test Network (FTN), and the National Board Certification (NBC) delivery system contractor in meeting the overall research and development objectives of the NBPTS.

- Conduct a field test of all assessment exercise packages and scoring procedures developed by the ADL, in cooperation with the NBPTS's FTN and TAG, including the training of administrators and assessors for the field test. This evolved over time to include a three-part pilot test—both local and national in scope—in place of a formal field test.

- Produce on schedule supporting resources and materials that are essential to the operational administration and scoring of complete assessment packages, including comprehensive informational materials for certification candidates, training materials for those who administer and score the assessments and feedback to candidates on their performances.

- Work collaboratively with NBPTS staff to discuss and review on a regular basis the plans for and progress of the assessment development and related efforts.

(NBPTS, 1994)

* The Technical Analysis Group was a group of advisors whose task was to oversee the psychometric soundness of the entire assessment process.

Figure 2

The Assessment Development Process

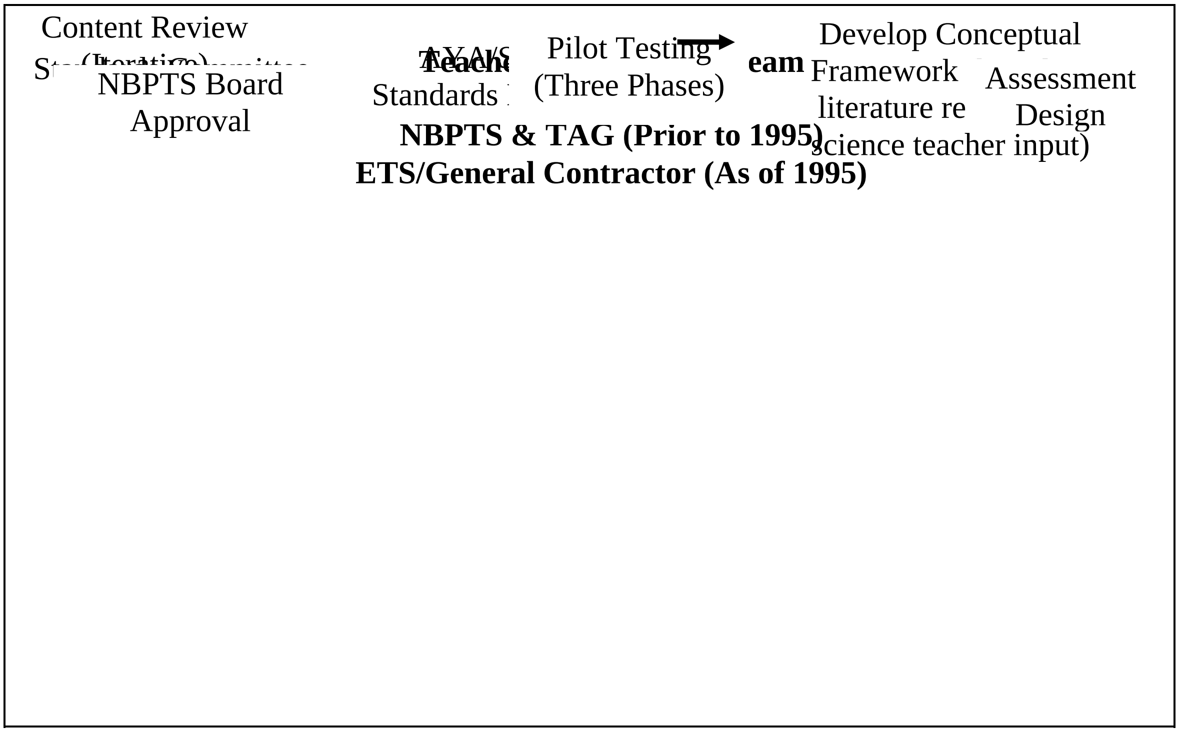


Figure 3

Comparison of the Evolving Portfolio Entries and Standards Coverage

The **Original 5** Portfolio Entries	The **Interim 9** Portfolio Entries	The **Final 6** Portfolio Entries
1. Teaching Toward the Development of Major Ideas in Science (Standards: I-Understanding Students, III-Instructional Resources, VI-Equitable Participation, VIII-Conceptual Understanding, IX-Contexts of Science)	1. Major Ideas (Standards: II-Knowledge of Science, VI-Equitable Participation, VIII-Conceptual Understanding, IX-Contexts of Science)	1. Teaching Major Ideas over Time (Standards: I-Understanding Students, II-Knowledge of Science, III-Instructional Resources, VI-Equitable Participation, VIII-Conceptual Understanding, IX-Contexts of Science, XIII-Reflection)

2. Assessing My Student's Work (Standards: II-Understanding Students, IV-Engagement, X-Assessment)	2. Assessing My Class (Standards: I-Understanding Students, IV-Engagement, X-Assessment) 3. Assessing Individual Students (Standards: I-Understanding Students, IV-Engagement, X-Assessment)	2. Assessing Student Work (Standards: I-Understanding Students, II-Knowledge of Science, VI-Equitable Participation, X-Assessment, XIII-Reflection)
3. Science Inquiry (Standards: I-Understanding Students, V-Learning Environment, VII-Science Inquiry)	4. Hands-on Investigation (Standards: III-Instructional Resources, IV-Engagement, V-Learning Environment, VII-Science Inquiry)	3. Hands-on Scientific Inquiry (Standards: II-Knowledge of Science, III-Instructional Resources, IV-Engagement, V-Learning Environment, VI-Equitable Participation, VII-Science Inquiry,

	5. Inquiry Through Discourse (Standards: IV-Engagement, V-Learning Environment, VII-Science Inquiry) 6. Real World Connections (Standards: IV-Engagement, V-Learning Environment, VII-Science Inquiry, IX-Contexts of Science)	VIII-Conceptual Understandings, XIII-Reflection) 4. Discussions about Science (Standards: II-Knowledge of Science, IV-Engagement, V-Learning Environment, VI-Equitable Participation, VII-Science Inquiry, VIII-Conceptual Understanding, IX-Contexts of Science, XIII-Reflection)

4. Professional Commitment (Standards: XI-Family & Community Outreach, XII-Collegiality & Leadership, XIII-Reflection)	7. Beyond the Classroom (Standard: XI-Family & Community Outreach) 8. Professional Commitment (Standard: XII-Collegiality & Leadership)	5. Documented Accomplishments I (Standard: XII-Collegiality & Leadership) 6. Documented Accomplishments II (Standard: XI-Family & Community Outreach)
5. Portfolio Reflections & Teaching Philosophy (Standard: XIII-Reflection)	9. Portfolio Cover Letter (Standard: XIII-Reflection)	

Figure 4

Narrative Overview of the Portfolio Entries

the portfolio entries

The five entries in the AYA/Science portfolio can be grouped in three categories based on the primary source of evidence: those based on student work, videotape clips, or your work outside the classroom. Descriptions of each entry are provided below based on this categorization.

Entries based primarily on student work samples:

The first essential source of evidence about your practice is student work: what do you ask your students to do, how do you interpret student responses, and what do you do with this information. There are many kinds of student work; these portfolio entries attempt to sample types that are important for science teachers of adolescents and young adults. In the Written Commentary for each entry, you will be asked to analyze each of the students' work samples and explain this work in the context of your teaching.

1. Teaching Major Ideas Over Time

In this entry you will demonstrate how you weave instructional activities together to promote students' understanding of one major idea in science. You will need to submit an **Overview of Instruction**, three **Activity Descriptions***,* each accompanied by two **Student Work Samples**, and a **Written Commentary** that collectively describe and illustrate your approach to teaching a major idea in science over time.

2. Assessing Student Work

In this entry you will demonstrate how you assess the progress of your students during a period of instruction. You will need to submit three Assessment Descriptions, each accompanied by two Student Work Samples, and a Written Commentary that collectively describe and illustrate your approach to assessment.

You must feature different classes in each of these two entries based on student work. In addition, you must select a different lesson or unit for each of the four classroom-based entries. For further guidance, refer to the Entry Tracking Form on page 11.

Entries based primarily on videotape clips:

There is no better evidence of what a teacher does than actual classroom practice. For this reason, video clips of practice in varying situations and circumstances are essential evidence for the accomplishments of teachers. These portfolio entries sample a teacher's classroom practice across different classes and across different content during the year. In addition, these video clips are designed to sample different kinds of instruction and classroom interactions. In the Written Commentary for each entry, you will be asked to analyze each of the video clips and explain the featured interactions in the context of your teaching.

3. Hands-on Scientific Inquiry

In this entry you will demonstrate your skill in engaging your students in hands-on scientific investigations that foster independent thinking. You will need to submit a **Written Commentary** and a **Videotape** that describe and illustrate your approach to engaging students in hands-on science to investigate a scientific concept. You will also submit **Instructional Artifacts** that clarify what is happening in the videotape.

4. Discussions about Science

In this entry you will show how you engage students in discussions that increase their understanding of science and make those discussions interesting, accessible and relevant. You will need to submit a **Written Commentary** and a **Videotape** that describe and illustrate how you engage students in discussing scientific concepts and solving problems in a scientific way. You will also submit **Instructional Artifacts** that clarify what is happening in the videotape.

You must feature different classes in each of these two entries based on videotape clips. In addition, you must select a different lesson or unit for each of the four classroom-based entries.

Each videotape entry must be accompanied by a photocopy of a government-issued photo ID such as a driver's license or school district ID. The photo ID should be enlarged to double its actual size, so that both your photo and your name are clearly visible. A sample of a photocopied government-issued photo ID appears at the end of this section.

Entry based primarily on work outside the classroom:

The third essential source of evidence about a teacher's practice reflects those aspects of teaching that do not occur in the classroom with students, but in a teacher's interactions with students' families, with the school and local communities, and with colleagues.

5. Documented Accomplishments

In this entry you will provide evidence of some of your professional work outside of the classroom. You will need to submit descriptions and documentation of those activities and accomplishments that illustrate your commitment to the families and communities of your students along with your contributions to the teaching profession.

Figure 5

Shell for AYA/Science Portfolio

Figure 5a

Excerpts from the Development Shell for the AYA/Science Portfolio

3. Instructions

3.1 Write the subheading, *what do I need to do?*

3.2 In a short paragraph describe what the candidate needs to submit. The last sentence should read in bold, *You must submit all of these products in order for your response to be scorable.*

3.3 Write the paragraph, *In the table on the next page, you will find a checklist of what you need to do to complete this entry. Following the table is a detailed description of how to make good choices as you plan and prepare your response. Specific requirements for each product are described in later*

Figure 5b

Excerpts from the AYA/Science Portfolio Entry 1

what do I need to do?

Submit three **Activity Descriptions**, each accompanied by two **Student Work Samples** (see illustration below); an **Overview of Instruction**; and a **Written Commentary** that collectively describe and illustrate your approach to teaching a major idea in science over time. **You must submit all of these products in order for your response to be scorable.**

In the table on the next page, you will find a checklist of what you need to do to complete this entry. Following the table is a detailed description of how to make good choices as you plan and prepare your response.

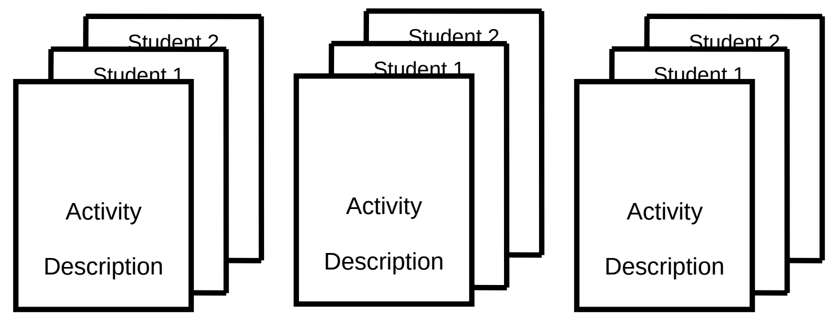
sections. 3.4 Provide a graphic that illustrates the products. _______________________page break____________________ 3.5 Provide a 3-column table with a checklist describing what the candidate needs to do to complete the entry. The columns should read, *BEFORE the period of instruction, DURING the period of instruction* and *AFTER the period of instruction*.	Specific requirements for each product are described in later sections. 

Figure 6

Description of Portfolio Entry 1: Teaching Major Ideas Over Time*

What is the nature of this entry?

In this entry you will demonstrate how you weave instructional activities together to promote students' understanding of one major idea in science. Through an overview of a period of instruction, selected activities, samples of student work, and a written commentary, you will provide evidence of how you teach over a period of time. This includes how you sequence and plan instructional activities, support your teaching with instructional resources, and illustrate the relevance of science to all of your students. This entry also asks that you reflect on your instruction and describe, analyze and evaluate how you promote student learning over a period of time.

* All examples from the AYA/Science assessment reflect the final version as developed by WestEd. Further revisions were implemented by the general contractor after delivery of the assessment to the NBPTS.

Figure 7

Evolution of "Instructional Context" Prompts

Early Version:		**Late Version:**
Provide information about the context of this class that is *relevant* to this entry. Remember that this response is intended to give information about the particular students in this particular entry. This might include such information as the subject matter of the class, the grouping of the students in the class, and individual details about the particular students who appear in the video. You might wish to add details about the "personality" of the particular class, or any other information that could help an assessor "see" your class.		What are the features of your teaching setting that are relevant to this entry? It is important that you help an assessor "see" your students and your class to understand better how it functions. In your description, include: the title, subject matter and size of the class; a description of your students' skills, knowledge and previous experiences that relate to the science you teach; the particular challenges this group of students represents; and any other realities of the social and physical teaching context (e.g., available resources, heterogeneity of the class, and ethnic and linguistic diversity) that are relevant to this entry.

Figure 8

Excerpt from Portfolio Entry 1 of the AYA/Science Assessment

Accomplished science teachers focus their curriculum to develop an in-depth understanding of the major ideas in science. This curriculum embodies accurate and coherent relationships between concepts, as well as connections to other scientific disciplines, students' personal experiences, "real life" applications, the larger social context, and the history of how scientific ideas and discoveries have evolved and matured. New ideas take on meaning when they are relevant, placed in larger contexts, and connected to previous experiences and knowledge …

Figure 9

Excerpts from the AYA/Science Standards

Standard II: Knowledge of Science

Accomplished science teachers have a broad and current knowledge of science and science education, along with an in-depth knowledge of one of the subfields of science, which they use to set important learning goals …

2) Fundamental Ideas of Science

Accomplished science teachers also possess a broad grasp of the fundamental laws, principles, theories, facts and ideas that constitute the body of scientific knowledge and of their associated vocabulary and terminology. Science is a collaborative social enterprise that builds on the achievements of previous generations. Exemplary teachers are conversant with the major conceptual paradigms that researchers have developed over the years in the core science disciplines and use that knowledge to inform their practice. The breadth of their knowledge base, organized by discipline, includes a firm understanding of the following aspects of science:

Physical Sciences

The basic properties of matter and principles governing its interactions; the forms energy takes, its transformation from one form to another, and its relationship to matter; motion and the principles that explain it; the nature of atoms and molecules, and the way in which atoms and molecules can be transformed into different arrangements of matter; and the forces that exist between and within objects and atoms...

Figure 10

Science Knowledge Domains

Physics

- Motions and forces
- Conservation of energy and increase in disorder
- Interactions of energy and matter
- Chemistry
- Structure of Atoms
- Structure and properties of matter
- Chemical reactions

Earth and Space

- Energy in the earth system
- Geochemical cycles
- Origin and evolution of the earth system
- Origin and evolution of the universe

Biology

- The cell
- The molecular basis of heredity
- Biological evolution
- The interdependence of organisms
- Matter, energy and organization in living systems
- The behavior of organisms
- The human organism

Common Aspects of Science

- I. Basic ideas in science
  - Systems, order and organization
  - Evidence, models and explanation

Change, constancy and measurement

Evolution and equilibrium

Form and function

II. Societal issues of science

IIa. Technology

Abilities of technological design

Understandings about science and technology

Science and technology in local, national and global challenges

IIb. Environment and Society

Personal and community health

Population growth

Natural resources

Environmental quality

Natural and human-induced hazards

Science as a human endeavor

Historical perspectives

III. Investigation skills

Abilities necessary to do scientific inquiry

Understandings about scientific inquiry

Nature of scientific knowledge

Figure 11

Depth (D) and Breadth (B) of the Knowledge Assessed by Emphasis Area

	Domains				
Teacher's Chosen Area of Emphasis	Physics	Chemistry	Earth and Space	Biology	Common Aspects of Science
Physics	**D**	B	B	B	*D*
Chemistry	B	*D*	B	B	**D**
Earth and Space	B	B	*D*	B	**D**
Biology	B	B	B	*D*	**D**

Figure 12
Final Assessment Center Structure

Block of Exercises	Activities	Type of Exercise	Completion time (minutes)[a]
1. Core concept I from major content area	Problem solving: data interpretation		30
(Depth)	Design an instructional strategy to teach a concept	essay	30
	Analyze student work based on science misconceptions		30
2. Facts and concepts in non-major areas	Answer three exercises in non-major content area I		30
(Breadth)	Answer three exercises in non-major content area II	brief essay	30
	Answer three exercises in non-major content		30

		area III		
3.	Integrated knowledge in major and non-major areas (Depth-Breadth)	Draw on knowledge from a non-major content area		30
		Discuss topic in major content area and relate to non-major content area	essay	30
		Content/context connections		30
4.	Core concept II from major content area (Depth)	Problem Solving: Use of Procedures		30
		Real-world connections in science	essay	30
		Issues of current relevance		30

NOTES

[1] WestEd, formerly Far West Laboratory for Educational Research and Development, is one of the original regional educational laboratories created by the United States Congress in 1966. WestEd is a not-for-profit agency committed to improving education through research, development and service.

[2] Note that several of the major players who were involved prior to 1995 were no longer involved after the NBPTS's reorganization and new players emerged.

[3] The U.S. Government was "shut down" due to a budget dispute in the U.S. Congress. This resulted in nine months of funding delays from the National Science Foundation.

[4] NBPTS identifies five criteria for assessment development abbreviated with the acronym "APPLE," which stands for Administrative feasibility, Professional acceptability, Public credibility, Legal defensibility, and Economic affordability. These were the criteria that all ADLs were instructed to use to develop assessments.

[5] The AYA/Science Standards Committee determined that it is possible for a candidate to submit a portfolio in one teaching area of emphasis but sign up for the Assessment Center in another area of emphasis.

CONTACT DETAILS

Author	Contact
Steve Schneider	Steve Schneider Director, Mathematics, Science, and Technology Program WestEd 730 Harrison Street San Francisco, California 94107 Phone: 650.381.6410 Fax: 650.381.6401 Email: sschnei@wested.org
Kirsten R. Daehler	Kirsten R. Daehler Co-director of the Science Cases for Teacher Learning Project WestEd 4200 Farm Hill Blvd Redwood City, CA 94061-1099 Phone: (650) 381-6402 Fax: (650) 381-6421 Email: kdaehle@WestEd
Kristin Hershbell	Associate Dean: Kristin Hershbell Charles Office of Research, Planning, and Grants City College of San Francisco 50 Phelan Ave. C-306 San Francisco, CA 94112 Phone: (415) 239-3677 Fax: (415) 239-3010 Email: kcharles@ccsf.edu
Jody McCarthy	
Jerome Shaw	Jerome Shaw Assistant Professor of Education University of California Santa Cruz Email: jmlshaw@ucsc.edu Office: 225 Crown College Phone: (831) 459-3291
Guillermo Solano-Flores	Guillermo Solano Flores Professor of Education School of Education, Room 205 University of Colorado at Boulder 249 UCB Boulder, CO 80309 Phone: 303-492-7695 Fax: 303-492-7090 E-mail: Guillermo.Solano@Colorado.Edu