Journal of Personnel Evaluation in Education 12:2 139–161, 1998

An Integrative Approach to Portfolio Evaluation for Teacher Licensure

PAMELA A. MOSS, AARON M. SCHUTZ, AND KATHLEEN M. COLLINS
University of Michigan, Ann Arbor, MI 48109-1259 pamoss@umich.edu

Abstract

The purpose of our overall research agenda is to develop and evaluate a methodology for the assessment of teachers in which experienced teachers, serving as judges, engage in dialogue to integrate multiple sources of evidence about a candidate to reach a sound conclusion. The project that provides the venue for this research agenda is the Interstate New Teacher Assessment and Support Consortium (INTASC), which is developing a portfolio assessment system to assist participating states in making a decision about teacher licensure. To develop the theoretical foundation necessary to support and evaluate such dialogic and integrative assessment practices, we turn, in part, to the tradition of philosophical hermeneutics, as a complement to psychometrics. In this article, we characterize and assess the processes in which judges, trying out an integrative approach to portfolio evaluation for the first time, engage as they collaboratively construct and document their conclusions, and we locate this work in the larger research agenda. The premise of this project, which is being carefully evaluated in the course of inquiry, is that these integrative practices cannot only lead to an epistemologically sound evaluation of teaching but also promote an ongoing professional dialogue of critical reflection on teaching practice.

Current calls for reform in the professional development and practice of teachers consistently highlight particular themes—that successful teachers integrate complex evidence about their students' learning and the context in which they work, that they engage in ongoing critical reflection about their own teaching practices, and that they work as members of active learning communities (e.g., Darling-Hammond, 1995; INTASC, 1992; Lieberman, 1990; Little, 1993; Lord, 1994; Meier, 1995; NBPTS, n.d.; National Commission on Teaching and America's Future, 1996; Richardson, 1990; Tyack & Cuban, 1995). Consistent with this reform effort is the movement toward performance-based licensing and certification of teachers and the preparation of accomplished teachers to serve as judges for these assessments (INTASC, 1995; NBPTS, 1994). These are important steps toward bringing assessment practices in line with the goals of reform.

As important as these steps are, there exists a substantial disjunction between the assessment practices teachers typically engage in as judges or readers in large-scale assessment and the goals of the current reform movement. While psychometric theory has advanced to accommodate the evaluation of complex performances (e.g., Cronbach, Linn, Brennan & Haertel, 1995; Wiley & Haertel, 1994; Mislevy, 1994), conventional procedures for evaluating these performance assessments typically have individual readers working independently, scoring one exercise at a time, blind to the candidate's performance on other exercises. To reach a decision about certification, these independent scores are algorithmically combined and compared to a cut score predetermined by a separate process. Even methods of setting performance standards typically ask judges to examine partially decontextualized information rather than collected performances for an individual candidate (Jaeger, 1996; Jaeger, Mullis, Bourque & Shakrani, 1995). Thus, in the actual evaluation process, conventional practices of assessment essentially preclude collaboration and critical reflection that integrates evidence from multiple sources. While these scoring practices represent sound and thoughtful work within the psychometric tradition, it is useful to step outside that tradition to question both the quality of information and the consequences—not just on those assessed but on all stakeholders in the assessment process—of evaluating the portfolios in this way. (See Darling-Hammond, 1995; Darling-Hammond & Millman, 1990; Good, 1996; Haertel, 1991; Tellez, 1996; and Haney, Madaus & Kreitzer, 1987, for reviews of current and past work in teacher assessment.)

The purpose of our overall research agenda is to develop and evaluate a methodology for the assessment of teachers in which experienced teachers, serving as judges, engage in dialogue to integrate multiple sources of evidence about a candidate to reach a sound conclusion. The premise of this project, which is being carefully evaluated in the course of our inquiry, is that these integrative practices cannot only lead to an epistemologically sound, perhaps sounder, evaluation of teaching but also promote an ongoing professional dialogue of critical reflection on teaching practice. Hence, the process of assessment can serve to contribute to the development of new forms of professional discourse, a crucial resource for teachers' learning (D. Ball, personal communication, April 1996; Lord, 1994).

Such collaborative and contextualized assessment practices have been used for consequential decisions in a few local contexts in both teacher education (e.g., Alverno College Faculty, 1994; Kimball & Hanley, 1995; Lyons, 1995) and K–12 education (Darling-Hammond, Ancess & Falk, 1995; Mabry, 1992; Meier, 1995). However, existing theory in educational measurement does not provide an adequate epistemological basis to support this promising work. Moreover, validity practices theorized and used locally, where there is extensive contextualized information about the candidate, may not be adequate to support large-scale high-stakes assessment, where information about the candidate is limited to evidence provided in the assessment. Until a credible theoretical foundation (and feasible empirical demonstration) is available, it is unlikely that collaborative and contextualized evaluation practices will be widely used in large-scale assessment contexts. And the educational community loses a productive resource for fostering the kind of critical, evidence-based dialogue about teaching practice that is widely viewed as crucial to educational reform.

In this article, we characterize and assess the processes in which judges, trying out an integrative approach to portfolio evaluation for the first time, engage as they collaboratively construct and document their conclusions. We also provide an overview and theoretical rationale for the larger research agenda. The project that provides the venue for this research agenda is the Performance Assessment Development Project (PADP) of the Interstate New Teacher Assessment and Support Consortium (INTASC) which, as part of its work, is developing a portfolio assessment system to assist participating states in making a decision about teacher licensure.

To develop the theoretical foundation necessary to support and evaluate such dialogic and integrative assessment practices, we turn, in part, to the tradition of philosophical hermeneutics. Unlike practices based in psychometric theory, in which judges typically engage in independent readings of isolated performances, in a more hermeneutic approach readers work together to construct a coherent interpretation, continually challenging and revising initial interpretations, until they account for all the available evidence about a candidate. Our validity research agenda draws on principles from both hermeneutics and psychometrics to provide a rigorous evaluation of the assessment system.

The paper is organized into five major sections. In the first section, we provide a selective overview of those aspects of hermeneutics that directly inform our work. In the second section, we provide an overview of the INTASC portfolio assessment project and the larger validity research agenda of which the study reported here is a part. In the third section, we describe the processes of reader preparation and portfolio evaluation as they occurred for the mathematics field test in the summer of 1996 when we collected our data. In the fourth section, we present results from our studies of portfolio readers' processes. Our emphasis in the data analysis has been to seek out and illustrate problems—to focus on what's not working the way we had anticipated—so that INTASC can improve the processes of preparation and evaluation. We conclude the article, in the fifth section, with a discussion of the implications of this work for the professional development of teachers and for educational assessment more broadly.

Hermeneutics as a Theoretical Basis for Integrative Assessment Practices

A promising theoretical direction for more contextualized and collaborative approaches to assessment can be found in the research tradition of hermeneutics. Like psychometrics, hermeneutics characterizes a general approach to the interpretation of human products, expressions, or actions. Also like psychometrics, hermeneutics provides means of combining information across multiple pieces of evidence and of dealing with the disabling biases that readers may bring. The differences between these disciplines lie in the ways in which the information is combined and readers' biases are addressed. Although hermeneutics is not a unitary tradition, most hermeneutic philosophers share a holistic and integrative approach to interpretation of human phenomena, which seeks to understand the whole in light of its parts, repeatedly testing interpretations against the available evidence, until each of the parts can be accounted for in a coherent interpretation of the whole (Bleicher, 1980; Ormiston & Schrift, 1990; Schmidt, 1995). This iterative process is often referred to as the hermeneutic circle.

The approach to hermeneutics on which we draw most heavily is based on the hermeneutic philosophy of Gadamer (1987). Here the hermeneutic circle can be characterized as representing a dual dialectic—one between the parts of the text and the whole, and one between the text and the reader's foreknowledge, preconceptions, or ''enabling'' prejudices. Gadamer argues that there is no knowledge without foreknowledge—without preconceptions or prejudices. ''The task is not to remove all such preconceptions, but to test them critically in the course of inquiry ... to make the all important distinction between blind prejudices and ‘justified … [or enabling] prejudices that are productive of knowledge’ ’’ (Bernstein, 1985, p. 128).

The process of testing preconceptions in the course of inquiry begins with the respectful assumption that the text, which may contain apparent holes or contradictions, is coherent and can inform us of something (Gadamer, 1987; Taylor, 1987). The dialogue between the text and the reader, and among readers, is guided by the intent to understand and learn from the text. Through this process, we raise to a conscious level those ‘‘prejudices which govern understanding’’ (Gadamer, 1987, p. 137) and enable them to evolve. A critical elaboration of philosophical hermeneutics, ‘‘depth hermeneutics’’ (Habermas, 1990), which is informed by critical theory, locates validity in the consensus among readers who approach one another as equals, self-conscious of the ways that different ideologies (or biases) may be constraining their interpretations. Of course, systematically distorted communication—influenced by social, political, or economic forces of which interpreters are unaware—can result in a ‘‘false consensus.’’ Here, some theorists of critical hermeneutics (e.g., Hoy and McCarthy, 1994; Kogler, 1996) highlight the role that those from outside the interpretive community can play—bringing an alternative perspective that illuminates the values and theories taken for granted by those within the interpretive community, so that they may be self-consciously considered.

From a hermeneutic perspective, the process through which general principles or standards are applied to particular cases is dialectical: rather than simply applying fixed principles to particular cases, the meaning of the case and the principles are codetermined. As Gadamer suggests (1987), the hermeneutical process used in making a legal judgment exemplifies the hermeneutical process as a whole. He argues that ‘‘the judge does not simply ‘apply’ fixed, determinate laws to particular situations. Rather the judge must interpret and appropriate precedents and law to each new, particular situation. It is by virtue of such considered judgment that the meaning of the law and the meaning of the particular case are codetermined’’ (p. 148).

Thus, hermeneutic philosophy points to an integrative and dialogic approach to assessment—where readers’ developing conceptions of competent performance are continually evaluated; where the meaning of the principles guiding the assessment is mediated by the contingencies of the cases to which they are applied; where interpretations are continually revised until they account for all the available evidence; and where the validity of the conclusion is warranted, in part, in the consensus among readers who are empowered to challenge one another’s developing interpretations in light of the case at hand.

Interstate New Teacher Assessment and Support Consortium (INTASC)

The project that provides the venue for this research agenda is the Performance Assessment Development Project (PADP) of the Interstate New Teacher Assessment and Support Consortium (INTASC). INTASC, a program of the Council of Chief State School Officers (CCSSO), was established in 1987 to enhance collaboration among states in promoting reform in the education, licensing, and professional development of teachers.

INTASC's mission is to provide a forum for the states to learn about and collaborate on the development of programs to enhance the preparation, licensing, and professional development of teachers. In conjunction with the assessment development project described below, INTASC is fostering systemic reform by developing policies and practices that shape teacher preparation, program review, and ongoing professional development. INTASC began its work by crafting model standards for beginning teaching. To ensure compatibility and continuity in a teacher's career development, INTASC used the framework of the National Board for Professional Teaching Standards (NBPTS) to construct its core principles (INTASC, 1992). More recently, INTASC has been developing discipline-specific standards and performance assessments in mathematics and English language arts to be used in licensing decisions.

The INTASC plans for the portfolio assessment reflect two equally important goals—one is to assist participating states in making a sound decision about licensure in conjunction with additional evidence about the candidate, and the other is to develop an assessment and support system that encourages an ongoing professional dialogue of critical reflection on teaching practice. In the evaluation process under development, two readers (or judges) collaboratively evaluate the entire set of performances contained in the portfolio using interpretive categories based on INTASC standards as a framework to guide the gathering and analysis of evidence. Consistent with hermeneutic principles, the goal is to construct a coherent interpretation based on the entire set of portfolio entries, continually challenging and revising initial interpretations until they account for all the available evidence. The evaluation process takes readers through a series of explicit analytic stages involving data reduction and integration, each of which is recorded for use at the next stage. The readers then engage in dialogue to arrive at a consensus decision on the candidates' level of performance with respect to the INTASC standards.

The Validity Research Agenda

The validity research agenda begins with the assumption that the INTASC general and discipline-specific standards provide the preliminary definition of the construct of teaching competence that the validity research agenda will elaborate. Preliminary logical analyses document the coherence among INTASC general and content-specific standards, the portfolio tasks and evaluation criteria, exemplars (evaluated portfolios), the literature on teaching knowledge and practice, and the self-reported practices of teachers. Additional logical analyses review the assessment materials for sensitivity to candidates from diverse social backgrounds and document the legal defensibility of this approach to assessment.

Plans for validity research draw on both psychometric and hermeneutic principles to provide rigorous standards of evaluation. For most sources of evidence, consistent with psychometric principles of validity, we develop a contrast that has a certain amount of independence built in. For instance, we contrast independent evaluations of the same portfolios by different reader pairs (as described below); independent evaluations of two portfolios from the same candidate; alternative means of obtaining evidence of the same candidates, including candidates' self-reports and case studies of candidates in their local context; and alternative means of evaluating the portfolios, including readings by discipline-specific scholars drawing on their own criteria of good teaching. And then, consistent with hermeneutic principles of validity, we reconcile any differences we observe dialectically. In that way, we consider how the independent means of evaluation work together to form a more complete whole and highlight what might be missing from the picture of teaching practice when the only evidence available is that provided in the planned assessment process. In the case of portfolio readings (as described below), we also trace the process in which readers engage to see how their theories of competent beginning teaching evolve—how their preconceptions are challenged by the evidence—as a result of their experience. And we trace the process in which candidates engage as they prepare their portfolios to see if their activities are consistent with the goals of the project and facilitative of their learning.

Concerns about equity cut across all of these sources of evidence. Wherever possible, we look to see whether there is evidence of differences in the validity of the assessment for candidates who differ with respect to ethnicity, gender, and social context of work. To help us address issues of fairness, we have constituted an independent panel that will review the results of these and other studies, including raw data such as samples of the transcribed dialogue and readers' notes, along with other written documentation. Among the issues to be addressed by the panel are whether the assessment (1) provides an appropriate range of opportunities for candidates to display expertise, (2) clearly explains what is valued, (3) allows for differing resource levels and teaching environments, and (4) recognizes differing theories about teaching consistent with professional consensus about sound practice. This research agenda and its participants serve, in part, the purposes called for in critical hermeneutics by bringing an outside perspective into dialogue with the focal interpretive community, so that disabling biases can be illuminated and enabled to evolve.

The Processes of Reader Preparation and Portfolio Evaluation

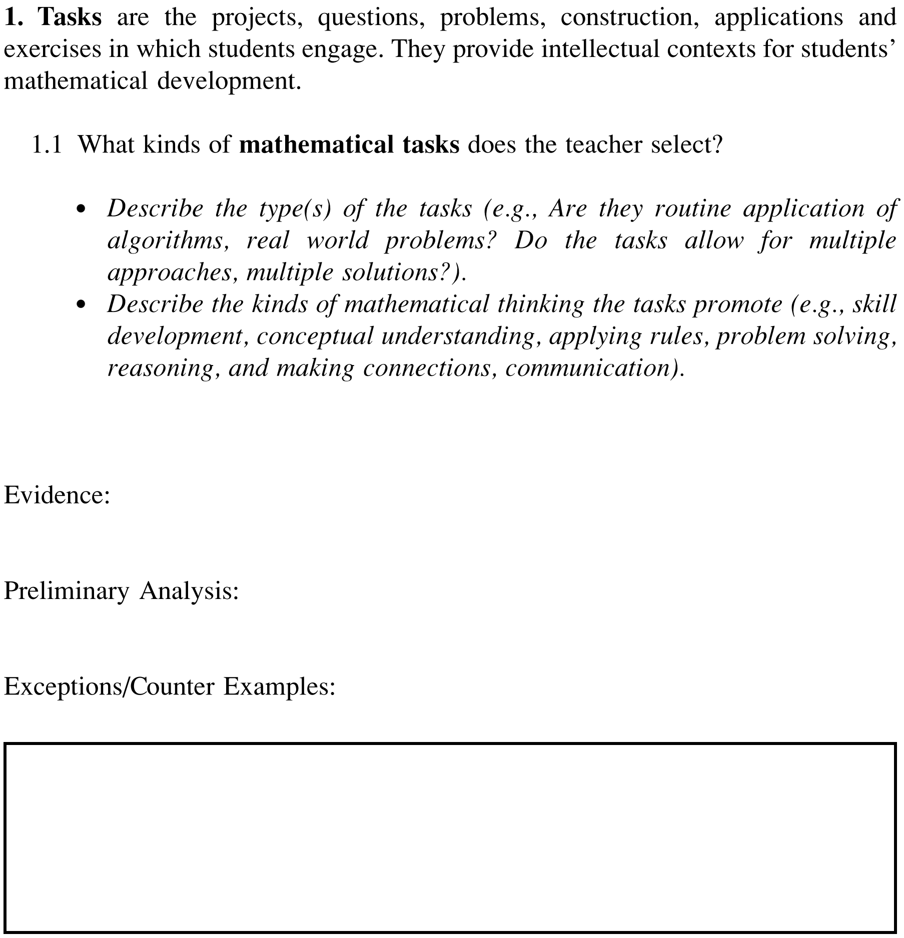
In this section, we characterize the processes of reader preparation and portfolio evaluation for the mathematics portfolio assessment field test as it occurred in August 1996. The readers involved in the summer portfolio evaluation had varying degrees of experience with INTASC and the Performance Assessment Development Project (PADP)—ranging from those who had participated in drafting the INTASC Mathematics Standards and Portfolio Handbook to those who joined us for the first time in August 1996. For all the readers, however, this was the first time using the PADP evaluation system to assign performance levels to portfolios.[1]

Reader preparation occurred over a five-day period, during which readers were introduced to the process of evaluating portfolios and given the opportunity to study portfolios presented as preliminary benchmarks for the different performance levels. The benchmarks had been selected earlier in the summer by a committee of twelve mathematics educators, seven of whom also served as evaluators in August.[2] As a result of discussions at the workshop, one of the benchmarks was reassigned to a different performance level. The reader preparation process was considerably briefer than is anticipated when the system is fully implemented. We anticipate at least ten days of preparation, over a one-year period, with a built-in process of certification for readers, who will commit to evaluating portfolios for at least three years.

The contents of the PADP portfolios are described in detail in a handbook sent to each candidate (INTASC, 1995). The handbook asks candidates for licensure to prepare an extensive portfolio with entries that document and reflect critically on their teaching practices. Candidates are encouraged to work with a mentor or other experienced colleague(s) as they prepare their portfolio. The portfolio entries are all organized around an eight- to twelve-hour set of lessons focusing on an important mathematical topic or idea. All entries ask for the candidates' commentary explaining the rationale for the choices they made and evaluating what actually occurred. In addition, candidates are asked to select three students from the class whose work they will trace and highlight through each of their portfolio entries. Specific entries include (1) a written commentary that sets the context for instruction in terms of the community, the class, the students, and the colleague with whom they will work and that describes the texts, technology, and other materials they use; (2) a set of plans for lessons and commentary that highlights the mathematics, types of tasks, discourse, environment, and analysis of learning typical of these lessons and that gives the rationale for these choices; (3) detailed descriptions of two featured lessons including videotapes, instructional materials, samples of student work, and commentary highlighting the rationale for their choices and their evaluation of what occurred; (4) a cumulative evaluation of student learning from these lessons, including a copy of the assessment and scoring criteria, samples of student work and feedback provided to those students, and a commentary on the assessment and its results; (5) a commentary reflecting the teacher's self-assessment and plans for professional development.

In the evaluation process, two readers worked together to evaluate a portfolio, integrating evidence from across the various performances represented, using interpretive categories based on INTASC standards as a framework to guide the gathering and analysis of evidence. The categories (mathematical tasks, mathematical discourse, learning environment, analysis of learning, analysis of teaching) were each elaborated with one to four guiding questions for readers to consider and the guiding questions were, in turn, illustrated with one to five bullets each highlighting aspects of the performance that might be considered.[3] Readers were expected to address all of the fourteen guiding questions but not all of the forty bullets (see figure 1 for an example).

Readers first worked through the portfolio alone, noting and recording evidence relevant to any of the interpretive categories wherever it occurred. Although readers had been given samples of notes, they were allowed to take notes in whatever way was most useful to them. Then readers worked together in pairs to prepare summary statements with supporting evidence for each guiding question, following the evaluation framework described above. The evaluation form on which readers formally recorded their evaluation was organized by interpretive category and guiding question. Underneath each guiding question and its associated bullets were places to record evidence, preliminary analysis, exceptions and counterexamples, and the final summary statement in response to the guiding question (see figure 1).



1. Tasks are the projects, questions, problems, construction, applications and exercises in which students engage. They provide intellectual contexts for students' mathematical development.

1.1 What kinds of **mathematical tasks** does the teacher select?

- *Describe the type(s) of the tasks (e.g., Are they routine application of algorithms, real world problems? Do the tasks allow for multiple approaches, multiple solutions?).*
- *Describe the kinds of mathematical thinking the tasks promote (e.g., skill development, conceptual understanding, applying rules, problem solving, reasoning, and making connections, communication).*

Evidence:

Preliminary Analysis:

Exceptions/Counter Examples:

Figure 1. Excerpt from ''Evaluation Form'' (Guiding Question 1.1 Under Tasks).

To assist readers in writing their summary statements, staff prepared an extensive list of ''sentence starters'' for each of the guiding questions to give readers a sense of how they might begin to characterize a portfolio performance.[4] It was not intended that readers would always use the sentence starters or that, when they did, they would duplicate them as is; rather, the sentence starters were intended only as examples.

In preparing their summary statements, readers were encouraged to consider all the evidence in the portfolio, looking across the videotapes, the student artifacts, and the teacher's plans and reflections; to actively seek counterexamples that challenged developing interpretations; and to value conflicting interpretations, reaching consensus if possible, and documenting differences where consensus could not be reached, continually testing developing interpretations until all the evidence has been considered.

After completing the evaluation forms, readers were asked to reach consensus on one of five overall levels of performance and to record the level, with no need for further written justification. To help them to remember the benchmarks that represented the different levels of performance, a one-page summary was prepared, contrasting the benchmarks according to the five interpretive categories from the evaluation framework.[5]

Prior to the actual portfolio evaluation, the lead developer for the mathematics assessment selected twenty-eight portfolios to be evaluated. They reflected a full range of performance levels (based on his initial skim), included all of the portfolios from persons of color that were complete, and were balanced in terms of gender, teaching level, and teaching context of the candidate. The sixteen readers were assigned to eight reader pairs who stayed together throughout the subsequent evaluation session.[6] Each reader pair evaluated seven portfolios. All portfolios were evaluated independently by two reader pairs. Reader pairs were grouped into four ''quads'' so that each quad saw the same seven portfolios. As to the order of portfolios evaluated within quad, the second, sixth, and seventh portfolios were always the same for each pair in a quad; the order of the other portfolios was systematically rotated. There was no process of certification for readers; all readers who participated in the training participated in the portfolio evaluation. Reader pairs were given four days to evaluate their seven portfolios with the option of using four consecutive days or taking a two-day break after the second day. All reader pairs completed their assignment of seven portfolios within four days. One pair completed all seven portfolios is three days, and most pairs had completed their sixth portfolio by the end of the third day.

Studies of Readers' Processes

As we indicated in the introduction, the goal for the research undertaken in summer 1996 was formative. Our purpose was to ferret out problems in the processes of reader preparation and portfolio evaluation, as implemented, so as to enable the assessment system to evolve. A number of data sources are available to us, only some of which have been used in the analyses reported below. The data sources available to us, in addition to the original portfolios, include (1) audio tapes of interviews with five of six readers who were new to PADP prior to the summer workshop, (2) audiotapes of all training sessions, all small-group work during the training, and all dialogues between readers as they evaluated the portfolios (although a few of these tapes are not usable due to technical problems, (3) audiotapes of individual exit interviews with all readers after they had evaluated their sixth portfolio, (4) completed evaluation forms from each reader pair for every portfolio they evaluated, and (5) copies of readers' individual notes, when notes were taken in a way that could be duplicated (for example, one reader used multicolored stickies to indicate evidence relevant to each category, which she reused with each portfolio). In each section below, we describe the data sources we drew on for the analysis.

Below we report our methodology, progress, and findings from analyses within and across reader pairs. To illustrate the potential of this approach, we initially trace the process of one of the stronger pairs of readers as they evaluate a portfolio. We complement these analyses with information from semistructured interviews with individual readers that occurred toward the end of their experience. Then, and at greatest length, we compare the written documentation and dialogue from a sample of portfolio evaluations where two pairs of readers independently evaluated the same portfolios. Here, we look specifically for *differences* in interpretation to understand how and why differences occur.

In determining what to look for in readers' dialogue and written documentation, we have drawn on principles from both psychometric and hermeneutic traditions. *Within reader pairs*, we are interested in (1) the extent to which readers' developing interpretations or preconceptions of a candidate's performance are being regularly challenged, elaborated, or conditioned (a) by the evidence available in the portfolio and (b) by the other member of the pair and (2) the extent to which readers' interpretations reflect comparison and integration of evidence from multiple sections of the portfolio and multiple types of evidence, including the candidate's own explanations as well as the artifacts and video observed. *Across reader pairs*, we are interested in differences in interpretation that arise when reader pairs, working independently, interpret, and evaluate the same portfolio. If we can locate and understand the genesis of differences in interpretation, we can make recommendations for revising reader preparation so as to foster more consistent interpretations of the portfolios.

Case Study of One Reader Pair

In this section, to show the process readers went through in arriving at a decision, we draw from the dialogue of one of the stronger reader pairs, summarizing and excerpting from over 130 pages of written and transcribed data, including transcripts of the readers' dialogue, transcripts of exit interviews with each reader, and written portfolio evaluation forms completed by the readers together.

These two readers, Christine and Elizabeth, are experienced teachers of mathematics who did not know each other prior to working for INTASC. Elizabeth taught most recently at a suburban middle school, and Christine teaches at a vocational high school. Elizabeth is white, and Christine is African–American. The portfolio the readers are evaluating here is that of a teacher who works in a suburban middle school.[7]

As they evaluated the portfolio together, Christine and Elizabeth followed the organizational structure of the evaluation form, discussing, in order, tasks, discourse, environment, analysis of learning, and analysis of teaching. This pair tended to begin a new section by reading each guiding question out loud, after which they engaged in a discussion with three discernible phases. First, they each cited evidence that they had written down individually, alone, before they met to discuss the portfolio. When one noted evidence that the other did not have on her list, the pair often returned to the source of the evidence, either the written portfolio or their written scriptings of the videos, or both, and discussed the additional evidence. In the second phase, the pair generally looked for and discussed counterevidence, seeking to challenge their reading of the initial evidence. The third phase involved writing a summary statement that reflected their joint interpretation of the evidence and counterevidence on the evaluation form. Typically the pair began writing this statement by first trying out several of the sentence starters before creating their own description.

Given space constraints, we have chosen to excerpt vignettes from our extended case study that illustrate these readers' regular practices of integrating evidence from different areas of the portfolio, seeking counterexamples to challenge their initial assumptions, pointing to evidence that the other might not have noticed, and engaging in evaluations that represented the complex and sometimes contradictory nature of the evidence contained in this portfolio.

In their discussion of mathematical tasks, the pair agreed that many of the tasks assigned by the beginning teacher had the *potential* to promote conceptual learning. However, they also concluded that the teacher did not utilize this potential in her actual implementation of the tasks. The readers cited extensive, specific evidence from both the video and candidate's written commentaries to support their assertions, referring in this case to ten different specific activities in the teacher's lesson logs. For example, at one point in their dialogue about the tasks section, Elizabeth read the final summary they had just constructed for first guiding question aloud to Christine. Immediately after reading their statement, Elizabeth appeared to realize that they had relied heavily on the sentence starters. ''Oh my gosh!'' she said, ''That's a canned answer!'' Christine agreed, and the pair then revisited each assertion they had made in this statement, gathering additional evidence from the portfolio and the video to test each one. Part of their initial statement read ''the teacher uses a variety of tasks, including hands-on use of manipulatives.'' When testing this aspect of the statement, Christine cited the teacher's use of chips as evidence and then asked, ''What other hands-on manipulatives did she have other than the blue chips?'' Elizabeth answered, ''She used the paper strips for the number line,'' and then added, ''She used graph paper.'' Christine then contributed a point from her observation of the video, ''She used the cards, the three-by-five cards, index cards .... They didn't work, but she used them.'' After this discussion, the pair concluded that they had enough evidence to support this aspect of their initial assertion and so retained it as it was written. This example demonstrates the readers' generally careful and extensive examination of the portfolio contents as they searched for evidence, as well as their integration of evidence from different sources.

In their final summary for the first guiding question about discourse, the readers initially wrote that, ''The teacher asks questions that lead to single answers.'' Elizabeth and Christine did not stop there, however, and searched for counterexamples to challenge this statement. Elizabeth stated, ''She really does ask questions that might be used for exploration. But she—'' Christine finished Elizabeth's sentence for her: ''She did not let the students ... explore. Now ... on the video ... I think she did ask some questions they *could* have explored.'' Elizabeth replied, ''Except she gave them time to write their own equations. Is that exploring?'' Christine responded, ''I don't think they were exploring. They were just applying the rule .... See, she has already given them a sample .... So they weren't exploring anything, they were just following her sample.'' The pair agreed on Christine's interpretation and as a result added a statement to their final summary: ''The teacher asks questions that lead to single answers and some that might be used for exploration.'' Readers struggled in this manner to accurately represent the complex nature of this teacher's performance.

As they reviewed the evidence for the analysis of learning section, Elizabeth stated, ''There is no indication that students received [the quizzes] back. There is no ... evidence of how they were used to inform instruction. Or if students ever saw 'em again.'' Christine offered a counterexample: ''Now she did involve the students in self assessment .... It was ... in the review.'' Elizabeth and Christine looked for the students' self-assessment in the portfolio, and finding it, Elizabeth exclaimed, ''Oh, you're right!'' Elizabeth added that she was ''worried, because ... I knew that we were getting so many negatives and I just felt that there was something more that was here.'' Here we see examples of the way in which Christine and Elizabeth challenged one another's assertions with counterexamples from the text.

When it came time for Elizabeth and Christine to give the portfolio a final integrative performance level, the readers moved to the descriptions of the five possible performance levels, each of which consisted of six different aspects of performance. These aspects were intended to encourage the readers to integrate information from across the five different areas of the evaluation framework.[8] As they worked to reach consensus on a final score for this portfolio, Elizabeth and Christine found two different issues for which the performance-level descriptions did not address the specific performance they had seen for this teacher. For example, in analysis of teaching, even though they agreed that the teacher did note specific areas to improve and that the teacher's statements were generally accurate, which would indicate a score of 3, the readers nonetheless scored the teacher at a 2 (''analysis of teaching is very general statements and may be inaccurate'') because, as Elizabeth stated, ''There has to be more than recognition. There has to be a plan. We don't have any plan.'' The readers then agreed that the criteria should be changed to capture the full range of teacher performances. Thus, Christine and Elizabeth attempted to evaluate this portfolio in its specificity, noting areas where the statements provided under each of the performance levels did not encompass the unique nature of this teacher's performance.

There are aspects of Elizabeth and Christine's effort to arrive at a final performance level that may prove problematic, however. First, the dialogue does not show readers looking back through the summaries they have constructed when they shift to discussing a performance level decision. Second, although readers were not instructed to score each aspect as a separate item, our analysis reveals that nearly all pairs, including Elizabeth and Christine, took this approach. We return to this issue below.

Interviews with Individual Readers

Exit interviews were conducted with readers individually for approximately 45 minutes after they had completed their sixth portfolio. They were asked initially to reflect on their process of evaluating the sixth portfolio and then to reflect more generally on the process of training and evaluation. We were particularly interested in any thoughts that did not appear on the summary sheet or that were not expressed in dialogue. Findings relevant to this article are summarized below.

Disagreements between readers. In the context of hermeneutics, disagreement serves an epistemic function. It is one of the vehicles through which challenges to developing interpretations—necessary to monitor and control bias—are manifest. And it is one indicator of equal participation by readers in the process. In the interviews, none of the readers indicated that they had any major disagreements about the performance level they gave to any of the portfolios they read together, and readers indicated that there were few disagreements on the summary statements. Analyses of reader dialogue transcripts, which are described below, tend to agree with candidates' self-reports that there were few significant disagreements when pairs sought consensus on a performance level. However, the transcripts also indicate that there were more, often fairly subtle, disagreements and negotiations throughout the process of summary construction than readers noted, suggesting that by the time the pairs reached the point where they needed to decide on a performance level, many of their disagreements had already been worked out.

Readers differing roles. While readers may assume different roles to facilitate efficient completion of the portfolio evaluation, it is important that these differences do not undermine their ability to participate coequally in the process of constructing an interpretation of the candidates' performance. While most reader pairs indicated that one reader predominantly took on the task of actually writing their summary statements, few indicated that this led to any significant inequality in how these statements were constructed. Researcher observations and analyses of the transcripts, however, indicate that the writing and speaking roles established by most of the pairs may have had a negative impact on participation of one of the members.

Prejudgments. Again, it is one of the primary principles of hermeneutics that preconceptions, prejudgments, or ''biases''[9] be illuminated and challenged so they can be self-consciously considered and either accepted as enabling (in our case, consistent with INTASC principles) or revised to the extent that they are disabling. When asked to explicitly identify issues that they would have liked to have attended to, but that the evaluation framework did not address or allow, readers reported a wide range of issues. These included: a) aspects of classroom management that were not addressed by the guide; b) concerns about candidates' reliance on the textbook rather than showing their own thinking; c) gut feelings about the preparation of students in a candidate's classroom or the relationship between the teacher and students; d) struggles with personal responses to candidates; e) difficulties in judging candidates when it was not clear how much support a candidate had been given; f) difficulties with judging between well-equipped and poorer schools with different populations of students; g) and issues about how the reader would have done the same lesson differently. One reader said, ''If we saw something we were concerned about, that wasn't on, we couldn't figure out what pigeon-hole to put it in. We put it in anyway.'' Other readers clearly felt constrained in their answers by the framework. It's clear from the written documentation and dialogue that in some cases these issues influenced their judgment.

Analyses Across Reader Pairs

In this section we focus on comparative analyses of the interpretations of reader pairs working independently with the same portfolios. The cross-pair analyses focus on written documentation from a systematic sample of thirteen of the twenty-eight portfolios evaluated by two reader pairs and on transcriptions of the dialogue from a systematic subsample of four portfolios (supplemented with excerpts of dialogue from additional portfolio evaluations). As indicated above, we focused our analysis on finding and explaining *differences* in interpretation.[10] In the paragraphs that follow, we characterize common recurring issues.

Potential problems with readers' written evaluations. We noted that not all of the bullets under each question were being fully addressed by the readers. While it was never the intent that each of the forty bullets would be addressed, there are some patterns emerging that may need to be addressed. For instance, ''how'' and ''in what way'' questions tended to be ignored or to be turned into general descriptions rather than explanations. Similarly, questions that require detailed analysis of evidence (such as ''In what ways do students rely on mathematical evidence and argument to address validity?'') tended to be avoided or reduced to simpler evaluative statements. In addition, we noted that the analysis reflected in the summary statements appeared superficial in many cases (although subsequent analysis of the dialogue suggest far more analysis is going on than is reflected in these forms). For instance, we noticed frequent repetition of words from the bullets or sentence starters without additional descriptions specific to the portfolios. In general, the evidence recorded on the summary form, in support of the summary statement, was difficult to interpret. Although occasional interpretive phrases were recorded here, more typically activities or products were simply listed.

The effect of tentative hypotheses on the interpretation of evidence. In theory, readers who have constructed a tentative hypothesis about a portfolio are supposed to search for counterevidence that might either confront or complicate their hypothesis as they attempt to combine all the available evidence into an integrative evaluation of the candidate's performance. However, analyses of the summaries written by the readers, as well as of their dialogues, point to a number of potentially serious limits to this dialectical movement between hypothesis and challenging evidence. These problems seem to fall into two related categories. First, once constructed, tentative (or often not so tentative) *hypotheses* about the nature of the teacher's performance can affect how subsequent evidence is interpreted. Instead of *confronting* their assumptions, readers sometimes seemed to interpret evidence in ways that supported the assumptions they happened to already hold. Second, and deeply intertwined with the first, once interpretations have been agreed on for one aspect of a teacher's performance, these interpretations may then become givens on which readers base subsequent interpretations. An *interpretation* of one aspect of a teacher's performance or one piece of evidence can cascade through the rest of the portfolio, affecting a pair's subsequent interpretations. For example, in one case, pair C decided that the students performed acceptably on their cumulative exam, while pair D decided the students failed the test badly. Once arrived at, this decision affected their later interpretations of the portfolio. Pair C agreed with the teacher's interpretation of student performance on the exam and later argued that the teacher understood the kinds of changes she needs to make in her classroom. Pair D argues that the teacher misunderstood the implications of students performance on the exam and later stated that she did not understand the kinds of changes she needed to make in her teaching.

Evaluative overtones. In other portfolios, while the reader pairs seemed largely to describe the same aspects of a teacher's performance, they used very different evaluative language to frame these aspects. For example, in one case, reader pair E described a portfolio as representing a routinized, teacher-directed classroom, writing in their summary for tasks that ''the teacher's presentation is limited to discussing procedures and explaining directions for tasks.'' Reader pair F, however, described the same aspect of the same portfolio with the statement ''Tasks were presented meaningfully with clear directions.'' Thus, pair E emphasized the ''limits'' of the teacher's presentation, while pair F described her presentation in more positive terms. This difference in evaluative language persisted through the pairs' evaluations of the entire portfolio.

Video versus textual evidence: ''The Teacher States.'' Readers are often called on to decide whether what a teacher says happened (or will happen) actually did (or will) happen because they don't actually see it happen in the materials they are given. In their written documentation, some readers qualified descriptions based solely on what the teachers wrote in their commentary with phrases like ''the teacher states ... ,'' whereas other readers simply reported that these activities occurred. There is evidence that some readers drew on a range of indirect evidence to establish whether to believe a teacher actually did what she said, and the video was often crucial. For example, when a teacher says she will do something in written commentary and actually does it in the videos, some readers tend to be more inclined to believe that this teacher does what she says she will in her commentaries for which they had no corresponding videos. If there is a disjunction between the videos and the commentary, readers appear to be less inclined to believe what the teacher says.

Difficulties interpreting videos and student artifacts. For a number of the matched reader pairs, there seemed to be differences in the ways they interpreted interactions on the video and written evidence from student artifacts. In some cases, this related to differences in the extent to which readers struggled to hear or decipher the evidence that was presented. Often it was difficult (if not impossible) to hear the interaction on the video, especially student-student interaction, and to distinguish teacher's evaluations from students' responses on the photocopied artifacts (homework, tests, and so on). Often, the evidence available was incomplete, as when the camera failed to capture a significant event or when the teachers had not included the criteria they used in evaluating student work. In such cases, there were differences in the extent to which readers drew higher-level inferences from the available evidence. In the case of the video, this problem of drawing well-warranted inferences may be intensified during the reader's dialogue, as (given time constraints) they tended to rely on their notes and memories rather than returning to the video itself.

The final evaluation process. The descriptors of the performance levels that were given to the readers to use in assigning a final integrative score to the portfolios had six different aspects for each level that were broken out on separate lines. Except in the case of portfolios that seemed, in the judgement of the pairs, to fall clearly into the lowest category (a 1), all reader pairs tended to follow the same final decision process. The readers first decided whether the teacher was closer to a 2 or a 4. They then scored each aspect separately. In one case, the readers explicitly gave the teacher the score that was given to the largest number of aspects. Other pairs appeared to attempt to integrate the scores from the aspects in a more complex manner.[11] The descriptions of the different performance levels did not map in any simple way onto the structure of the evaluation framework. This was done to encourage the readers to combine their summaries into an integrative reading of the teacher's performance. In some cases, however, this may have created a problematic gap: issues that were not addressed by the readers as they constructed their summary statements were sometimes elicited by the performance levels, and issues that seemed important in the summary statements seemed sometimes to be lost as the pairs move to the performance levels.

Summary and Conclusions

Given the formative purpose of these studies, our conclusions are framed in terms of issues that need to be addressed. In the work subsequent to that described in this article, we have begun exploring a range of possible solutions. However, we have learned that changes in one area often reverberate in unexpected ways through the entire evaluation process. Thus, this article aims not at solutions but at framing some of the issues that others who are considering similar approaches must grapple with in order to achieve a fair and valid system.

In framing these issues and challenges, we focus on the principles emphasized in philosophical and critical hermeneutics about the habits of mind and practice that lead to sound interpretations. We have considered, as well, the principle emphasized in psychometrics, of consistency among independent readings. Further, we have considered the principles, emphasized in multiple traditions, of in-depth analysis and public accounting of evidence supporting conclusions so that others may audit and evaluate these conclusions for themselves.

Developing and Recording a Comprehensive Trail of Evidence Supporting Conclusions

In this section, we consider two interrelated concerns—the extent to which readers engaged in in-depth analysis of the evidence contained in each portfolio and the extent to which they are able to represent their analysis so that others may review it. As we have noted, our readings of the written documentation suggest that it will not, in general, enable a third party to see whether what is recorded supports the performance level decision. Although a comprehensive written record may be a reassuring indicator of an in-depth analysis, it is not a necessary condition. Experienced readers may well be able to reach a carefully reasoned decision based on notes that are not accessible to outsiders. And clearly, producing such an accessible audit trail will add time to the process of portfolio evaluation. At some point, INTASC staff, research consultants, and attorneys will need to consider the purposes this written documentation (notes, evidence records, interpretive summaries) is designed to serve, including the extent to which it should be accessible for audit by a third party.

The tendency of readers to write interpretive summaries that drew heavily on sentence starters or words from the guiding questions raised concerns (partially ameliorated by listening to the dialogue) about the depth with which readers were considering the contingencies of each portfolio. Consistent with this concern, many of the reader pairs we studied appeared to move to a decision about a performance level without explicitly reviewing their interpretive summaries. This raises concerns that the performance-level decision may be based on selective recollection rather than on a comprehensive weighing of available evidence. We also noted instances where readers were simply treating the benchmark summaries as elaborations of the performance-level descriptions, comparing their portfolios point by point to the benchmark summaries as they did with the performance levels, rather than treating the benchmarks as complex illustrations of one kind of performance often only predominantly consistent with that level. These are all issues that can be explicitly addressed in reader preparation.

In some cases, the portfolio, as prepared by the candidate, did not provide adequate evidence for readers to reach well-supported conclusions. Here, we noted problems in reading student artifacts and hearing and seeing videotapes. We also noted that in some cases the only evidence available to readers relevant to an important issue was the candidate's commentary. Clearly we must encourage and support candidates to both develop videotapes and artifacts that can be, to the extent possible, heard, seen, read, and understood, and to provide evidence, through videotapes and artifacts, to support the claims they make in their commentaries, particularly with respect to issues that weigh heavily in the determination of performance levels, so that readers can draw on multiple sources of evidence in supporting their conclusions.

Challenging Developing Interpretations, Seeking Coherence, and Empowering Equal Participation

In this section, we consider evidence relevant to these three interrelated principles of hermeneutics—regularly challenging developing interpretations with evidence from the portfolio and perspectives of other readers, seeking coherence among the available sources of evidence, and empowering readers to participate equally in the construction of the interpretation. As evidence (and common sense) suggest, readers do bring values and perspectives about teaching that are not covered in the evaluation framework—some of which may conflict with the framework, some of which may not be relevant (or not appropriately brought to bear) given the evidence contained in the portfolio, and some of which may reflect professional knowledge that enhances the evaluation. It will be useful to find ways to illuminate these perspectives so that they may be self-consciously considered and discussed. As further evidenced in the dialogue and interviews, individual readers in some pairs tended to take on differentiated roles that appeared, in some cases, to allow one reader's perspective to dominate. In addition, the evidence from readers' dialogues and their own self-reports regarding the absence of major disagreements indicated that there may be insufficient opportunity for readers either to develop their own perspectives *about the portfolio* before engaging in dialogue or for readers to maintain important disagreements during the process of summary writing. Instances of cascading interpretations, seeking confirmatory evidence for tentative hypotheses, and coloring interpretations with evaluative overtones suggest that some readers may need additional support in engaging comfortably with and eliciting challenges to their developing interpretations.

Consistency Among Independent Readings

In the sections where we compared independent readings of the same portfolios, we noted a number of instances of differing interpretations. In addition, different pairs often addressed different aspects of the guiding questions, potentially exacerbating differences in interpretation. While we assume that enhanced consistency at the level of the summary statements will enhance consistency of the decision at the performance level, we observed portfolios where such differences did not, in fact, result in different performance level decisions. Without additional evidence and analysis, it would be impossible to draw strong inferences about the relationship between differences on responses to selected guiding questions and the overall performance levels assigned. Providing readers with a wider array of benchmark performances at each level, particularly those that illustrate frequently arising issues, may help resolve some of the differences in interpretation.

In acknowledging the value of consistency among independent interpretations we also want to caution that lack of consistency is not necessarily an indicator of an unsound interpretation or of a flaw in the assessment system. There are a number of issues that have arisen on which readers, fully acculturated into the INTASC principles and capable of applying them consistently, might thoughtfully disagree: for instance, how to count otherwise worthwhile tasks that are known to have been drawn directly from the textbook, whether the ability level of the class (as characterized by the teacher) should be taken into account in evaluating the quality of instruction that might otherwise be judged fully consistent with a high performance level, or whether knowledge of limited resources or district policies inconsistent with INTASC principles should influence the conclusion.

And so low reliability, conventionally defined, from an initial reading, may simply indicate that this is a portfolio that needs additional attention because it raises a set of issues that the interpretive community has not yet considered. In the current assessment context, it will of course be necessary to document the extent to which, ultimately, it doesn't matter who the readers are that initially evaluate the portfolio. But the test of the coherence in the system is not consistency in these initial evaluations but rather consistency in the outcome for a portfolio that has received the additional attention and debate that its contingencies require. Over time, readers could begin to build up a repertoire of experience and dialogue around these kinds of issues so that the tendency to resolve them in similar ways may evolve along with the ability to recognize issues that need wider discussion.

One could, of course, work to minimize the complexity in the material that readers are reading, either by moving toward smaller segments of the portfolio or further constraining the kinds of practices that teachers can illustrate. That is the direction in which conventional psychometric practices might push us. However, that does not make these issues go away, it simply puts them to work behind our backs. A strong system is one that can document and take advantage of such moments of difference.

It could well be that the outcome of such differences in interpretation involves a rethinking of the criteria and values involved in the INTASC standards themselves. Evolving (and revolving) perspectives on curriculum and pedagogy over the past few decades and current debates among thoughtful proponents of alternative perspectives make it clear that there is no ideal vision of sound teaching that can be more and more closely approximated by better and better assessment practices. There are only contexts, more or less encompassing, that allow values and the theories and practices in which they are implied to coalesce for a given time. Standards will need to be regularly reconsidered in light of new perspectives and supporting evidence about what constitutes sound teaching practice. Again, a vital assessment system, it seems, is one that accepts this circumstance and orients itself to reflect critically on its experience so as to evolve in productive ways.

Implications for Reform

As we suggested at the beginning, our premise is that this integrative approach to portfolio evaluation not only can lead to an epistemologically sound evaluation of teaching but also can promote an ongoing professional dialogue of critical reflection on teaching practice. The set of studies reported here have focused on helping us improve the epistemological aspects of the assessment system. Readers' responses to questions we asked during their exit interviews provide some preliminary indication, from their own perspective, of the impact of this brief experience on their own professional development. While all readers found the process to be very hard work, to a person, they each reported having learned something valuable that they will take back to their school communities. Benefits they mentioned included using the evaluation framework to evaluate and improve their own teaching; discovering new ideas for classroom practice; understanding the value of performance based assessment for professional development, particularly with beginning teachers; and feeling empowered to begin discussion with colleagues based on the principles reflected in the evaluation framework. As we begin to trace readers' developing capabilities over multiple opportunities for portfolio evaluation, and to trace their own descriptions of how their practices and perspectives have changed across these opportunities, we will be better able to evaluate the potential of this project for widespread professional development, not just for beginning teachers but for their experienced colleagues as well.

Beyond the potential impact on candidates and readers, this kind of work has the potential to assist the community of teachers and teacher educators in developing and critically evaluating empirically-based performance standards for beginning teachers through a process that is meaningful and accessible to all members of the educational community. Unlike conventional practices of standard setting, where judgments about teaching are based on decontextualized pieces of information statistically combined to determine a standard, the performance standards can be created through extended dialogue among teachers and teacher educators, grounded in concrete, complex, and contextualized examples of teaching practice. As the initial setting of standards is publicly illustrated and articulated in annotated exemplars, a wider opportunity for critical review and dialogue about teaching practice is fostered. And through this process, standards of professional practice remain vital, available for ongoing review, and open to productive evolution as new exemplars are added to illustrate evolving conceptions of sound practice.

Acknowledgments

The studies described here are based on work supported by the Spencer Foundation and the Interstate New Teacher Assessment and Support Consortium (INTASC) Performance Assessment Development Project (PADP) of the Council of Chief State School Officers (CCSSO). The section on hermeneutics was developed while the senior author was working under a grant from the National Academy of Education/Spencer Postdoctoral Fellowship Program. Staff and consultants for the PADP have contributed, in varying ways, to all aspects of this work. We are grateful to Ray Pecheone, Bill Thompson, Penny Pence, Jean Miller, Linda Wurzbach, Allison Kaye, Laurie Somers, Deborah Ball, Mary Diez, Jim Gee, Anne Gere, Bob Linn, Diana Pullin, and Mike Rebell. Opinions, findings, conclusions, and recommendations expressed are those of the authors and do not necessarily reflect the views of CCSSO, INTASC, the PADP, or the states participating in the project. We are grateful to Ed Wall and Russ Jacobs for invaluable assistance with data analysis.

Notes

1. The pool of sixteen readers included eleven women and five men. There were four African–Americans and twelve whites. Eleven readers were working in secondary schools (representing a range of teaching contexts), two had moved from the secondary school context to consulting or administrative positions, and three worked in postsecondary schools.
2. The larger validity research agenda calls for a series of additional steps through which the benchmarks will receive wider professional review.
3. This is a process similar to the one that Delandshere and Petrosky (1994) developed in their early work in

teacher portfolio assessment, although their guiding questions and interpretive summaries focus on one exercise at a time, whereas ours encompass the entire portfolio.

4. For instance, for the first guiding question under tasks, readers were given twenty phrases as examples of the ways in which they might start their own sentences. Examples included ''The teacher uses tasks that require students to make conjectures . . . '' or ''The teacher selects tasks that are primarily applications of formulas . . . ''
5. When this system is operationalized, we anticipate there will be regular confirmatory readings examining the extent to which readers' interpretive summaries in fact support their score decisions, as well as independent rescoring of a random sample of portfolios to monitor reliability. In addition, we anticipate that for portfolios where readers could not reach a clear consensus about a pass-fail decision or where a score results in the candidates' inability to obtain a license, an additional review will be given. As yet, the specific criteria for determining which portfolios need an additional review and the specific processes through which that review will occur have not yet been determined. Candidates will also have the opportunity to appeal a decision.
6. In considering appropriate assignments to pairs, the lead developer consulted with INTASC staff and members of the research team who were present. Issues considered in pairing readers included differing contexts of teaching experience, differing levels of experience with the PADP, work styles that appeared to complement one another, and whether the readers had chosen to work straight through the four days of evaluation or to take the two-day break.
7. Additional details about the candidate and her classroom have been omitted to maintain confidentiality.
8. The performance level for the score of 2 on which they finally agreed reads as follows: ''Students primarily develop a procedural understanding of mathematics. They learn how to solve routine mathematical problems. Students use oral or written discourse to explain their thinking. The mathematics is primarily taught in one way, with little or no adjustment for the different ways in which students learn. Student work is corrected. The teacher's analysis of teaching is limited to very general statements and may be inaccurate.''
9. In the context of psychometrics, *bias* is an unequivocally negative term—something to be avoided wherever possible. As indicated in the previous section, in the context of hermeneutics, bias (synonymous with foreknowledge or preconceptions) cannot be avoided because it makes understanding possible. Rather, enabling biases must be distinguished from those that are disabling so that the disabling biases can evolve. To what extent these different uses of the term *bias* reflect differences in terminology or fundamental differences about the nature of social reality is beyond the scope of this article. Both traditions, however, recognize the crucial importance of managing biases that undermine the soundness of an interpretation.
10. As the above paragraphs imply, it was not our purpose in this set of studies to document consistency among readers or to estimate interreaderpair reliability. While it is possible, given our design, to calculate an estimate of interreaderpair reliability, such an estimate is difficult to interpret meaningfully. The preparation for readers was considerably briefer than intended when the system is operationalized (five consecutive days versus an anticipated ten days spread over one year), and there was no attempt to distinguish fully prepared readers from among the pool of readers (all readers who attended the workshop, including those completely new to the PADP evaluated portfolios). When the system is operationalized, the most important reader reliability estimate will focus on decision consistency for the pass-fail decision. Since there was no cut point determined for the performance levels used last summer, it is not possible to report reliability in terms of decision consistency for the pass-fail decision. For those who are interested, we offer the following information. Twenty-eight portfolios were evaluated, independently, by two pairs of readers, who produced interpretive summaries and performance levels (with five possible levels) for each portfolio. For twenty-four of the twenty-eight portfolios (86 per cent), reader pairs reached agreement on the performance level within one level (adjacent agreement); for eleven portfolios (39 per cent) reader pairs reached exact agreement on the performance level.
11. When readers had difficulty deciding on a performance level just from the aspects, they would then usually move to the summaries of the different benchmark performances provided to them and attempt to rank the teacher in terms of the categories and descriptions represented there. Many readers appeared to be comparing, question by question, the *summaries* of the benchmark portfolios instead of the benchmark performances as a complex and integrative whole. This is problematic, since the complex nature of

performances in the portfolios usually means that specific aspects of a candidate's performance will fall above or below the performance level the portfolio is given as a whole.

References

Alverno College Faculty. (1994). *Student assessment-as-learning at Alverno College*. Milwaukee, WI: Alverno College Institute.

Bernstein, R.J. (1985). *Beyond objectivism and relativism: Science, hermeneutics, and praxis*. Philadelphia: University of Pennsylvania Press.

Bleicher, J. (1980). *Contemporary hermeneutics: Hermeneutics as method, philosophy, and critique*. London: Routledge and Kegan Paul.

Cronbach, L.J., Linn, R.L., Brennan, R.L., & Haertel, E. (1995). Generalizability analysis for educational assessments. Los Angeles: UCLA Center for the Study of Evaluation and the National Center for Research on Evaluation, Standards, and Student Testing.

Darling-Hammond, L. (1995). *A license to teach: Building a profession for twenty-first-century schools*. Boulder, COL: Westview Press.

Darling-Hammond, L., Ancess, J., & Falk, B. (1995). *Authentic assessment in action: Studies of schools and students at work*. New York: Teachers College Press.

Darling-Hammond, L., & Millman, J. (1990). *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers*. Beverly Hills, CA: Sage.

Delandshere, G., & Petrosky, A.R. (1994). Capturing teachers' knowledge: Performance assessment a) and post-structuralist epistemology, b) from a post-structuralist perceptive, c) and post-structuralism, d) none of the above. *Educational Researcher*, *23*(5), 11–18.

Gadamer, H.G. (1987). The problem of historical consciousness. In P. Rabinow & W.M. Sullivan (eds.), *Interpretive social science: A second look* (pp. 82–140). Berkeley: University of California Press. First published in 1963.

Good, T.L. (1996). Teaching effects and teacher evaluation. In J. Sikula (ed.), *Handbook of research on teacher education* (2nd ed.) (pp. 617–665). New York: Simon & Schuster Macmillan.

Habermas, J. (1990). The hermeneutic claim to universality. In G.L. Ormiston & A.D. Schrift (eds.), *The hermeneutic tradition: From Ast to Ricoeur* (pp. 245–272). Albany: SUNY Press. First published in 1980.

Haertel, E.H. (1991). New forms of teacher assessment. In G. Grant (ed.), *Review of Research in Education*, *17*, 3–29.

Haney, W., Madaus, G., & Kreitzer, A. (1987). Charms talismanic: Testing teachers for the improvement of American education. In E.Z. Rothkopf (ed.), *Review of Research in Education*, *14* (pp. 169–238). Washington, DC: American Educational Research Association.

Hoy, D.C., & McCarthy, T. (1994). *Critical theory.* Oxford: Blackwell.

Interstate New Teacher Assessment and Support Consortium (INTASC). (1992). *Model standards for beginning teacher licensing and development: A resource for state dialogue*. Washington, DC: Author and Council of Chief State School Officers.

Interstate New Teacher Assessment and Support Consortium. (INTASC). (1995). *Next steps: Moving towards performance-based licensing in teaching*. Washington, DC: Author and Council of Chief State School Officers.

Jaeger, R.M. (1995). On the cognitive construction of standard-setting judgments: The case of configural scoring. In *Proceedings of the Joint Conference on Standard Setting in Large Scale Assessment*. Washington, DC: National Assessment Governing Board and the National Center for Education Statistics.

Jaeger, R.M., Mullis, I.V.S., Bourque, M., & Shakrani, S. (1995). Setting performance standards for performance assessments: Some fundamental issues, current practice, and technical dilemmas. In G. Phillips (ed.), *Technical Issues in Performance Assessment*. Washington, DC: Department of Education.

Kimball, W.H., & Hanley, S. (1995). Fair and defensible judging of teacher performance and understanding for certification and teacher improvement. Manuscript, University of Southern Maine.

Kogler, H. (1996). *The power of dialogue: Critical hermeneutics after Gadamer and Foucault*. Cambridge: MIT Press.

Lieberman, A. (1990). Introduction. In A. Lieberman (ed.), *The changing contexts of teaching* (pp. 1–10). Chicago: National Society for the Study of Education.

Lieberman, A. (1995). Restructuring schools: The dynamics of changing practice, structure, and culture. In A. Lieberman (ed.), *The work of restructuring schools: Building from the ground up* (pp. 1–17). New York: Teachers College Press.

Little, Judith Warren. (1993). Teachers' professional development in a climate of educational reform. *Educational Evaluation and Policy Analysis*, *15*(2), 129–151.

Lord, B. (1994). Teacher's professional development: Critical colleagueship and the role of professional communities. In N. Cobb (ed.), *The future of education: Perspectives on national standards in America* (pp. 175–204). New York: College Entrance Examination Board.

Lyons, N., & Faculty of the University of Southern Maine's Extended Teacher Education Program. (1995). Which standards? What performance? For what vision of teaching and learning? Manuscript, University of Southern Maine.

Mabry, L. (1992). *Alternative assessment in an American high school*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Meier, D. (1995). *The power of their ideas: Lessons for America from a small school in Harlem*. Boston: Beacon Press.

Mislevy, R.J. (1994). Evidence and inference in educational assessment. *Psychometrika*, *59*(4), 439–483.

National Board for Professional Teaching Standards. (n.d.). *What teachers should know and be able to do*. Southfield, MI: NBPTS.

National Board for Professional Teaching Standards. (1994). *Report to the U.S. Senate Committee on Labor and Human Resources, and the U.S. House of Representatives Committee on Education and Labor*. Detroit, MI: Author.

National Commission on Teaching and America's Future. (1996). *What matters most: Teaching for America's future*. Author: Washington, DC.

Ormiston, G.L., & Schrift, A.D. (eds.). (1990). *The hermeneutic tradition: From Ast to Ricoeur*. Albany: SUNY Press.

Richardson, Virginia. (1990). Significant and worthwhile change in teaching practice. *Educational Researcher*, *18*(7), 10–18.

Schmidt, L.K. (ed.). (1995). *The specter of relativism: Truth, dialogue, and phronesis in philosophical hermeneutics*. Evanston, IL: Northwestern University Press.

Shulman, L.S. (1987). Assessment for teaching: An initiative for the profession. *Phi Delta Kappan*, *69*, 38–44.

Taylor, C. (1987). Interpretation and the Sciences of Man. In P. Rabinow & W.M. Sullivan (eds.), *Interpretive social science: a second look* (pp. 33–81). Berkeley: University of California Press. First published in 1967.

Tellez, Kip. (1996). Authentic Assessment. In J. Sikula (ed.), *Handbook of research on teacher education* (2nd ed.) (pp. 704–721). New York: Simon & Schuster Macmillan.

Tyack, D., & Cuban, L. (1995). *Tinkering toward utopia: A century of public school reform*. Cambridge, MA: Harvard University Press.

Wiley, D.E., & Haertel, E.H. (1994). Extended assessment tasks: Purposes, definitions, scoring, and accuracy. In R. Mitchell & M. Kane (eds.), *Implementing performance assessment: Promises, problems, and challenges*. Washington, DC: Pelavin Associates.