7-20-2004

Education Policy Analysis Archives 12/32

Arizona State University

University of South Florida

EDUCATION POLICY ANALYSIS ARCHIVES

A peer-reviewed scholarly journal
Editor: Gene V Glass
College of Education
Arizona State University

Volume 12 Number 32 **July 20, 2004** **ISSN 1068-2341**

Interrogating the Generalizability of Portfolio Assessments of Beginning Teachers: A Qualitative Study

Pamela A. Moss
LeeAnn M. Sutherland
Laura Haniford
Renee Miller
David Johnson
University of Michigan

Pamela K. Geist
Denver, Colorado

Stephen M. Koziol, Jr.
University of Maryland

Jon R. Star
Michigan State University

Raymond L. Pecheone
Stanford University

Citation: Moss P.A., Sutherland, L.M., Haniford, L., Miller, R., Johnson, D., Geist, P.K., Koziol, S.M., Star, J.R., Pecheone, R.L., (2004, July 20). Interrogating the generalizability of portfolio assessments of beginning teachers: A qualitative study, *Education Policy Analysis Archives, 12*(32). Retrieved [Date] from http://epaa.asu.edu/epaa/v12n32/.

Abstract
This qualitative study is intended to illuminate factors that affect the generalizability of portfolio assessments of beginning teachers. By generalizability, we refer here to the extent to which the portfolio assessment supports generalizations from the particular evidence reflected in the portfolio to the conception of competent teaching reflected in the standards on which the assessment is based. Or, more practically, “The key question is, ‘How likely is it that this finding would be reversed or substantially altered if a second, independent assessment of the same kind were made?’” (Cronbach, Linn, Brennan, and Haertel, 1997, p. 1). In addressing this question, we draw on two kinds of evidence that are rarely available: comparisons of two different portfolios completed by the same teacher in the same year and comparisons between a portfolio and a multi-day case study (observation and interview completed shortly after portfolio submission) intended to parallel the evidence called for in the portfolio assessment. Our formative goal is to illuminate issues that assessment developers and users can take into account in designing assessment systems and appropriately limiting score interpretations. (Note 1)

Introduction

A growing number of states are using some form of standardized assessment to assist in the licensure decisions about beginning teachers. Among the 42 states requiring such tests in 2000, the most widely used were paper-and-pencil tests assessing varied combinations of basic skills, content knowledge, or pedagogical knowledge (NRC, 2001b). The National Research Council's "Committee on Assessment and Teacher Quality" concluded that "paper and pencil tests provide only some of the information needed to evaluate the competencies of teacher candidates" (NRC, 2001b, p. 69). The committee called for additional research into the development of licensure systems that include assessment of teaching performance. As evidenced in the work of the National Board for Professional Teaching Standards (NBPTS), portfolio assessment provides one credible means for the large-scale high-stakes assessment of teaching performance. The Interstate New Teacher Assessment and Support Consortium (INTASC) is building on the pioneering work of the NBPTS to develop subject-specific portfolio assessments of beginning teachers. Their work provides the basis for this study.

This qualitative study is intended to illuminate the factors that affect the generalizability of this portfolio assessment of beginning teachers. By generalizability, we refer here to the extent to which the portfolio assessment supports generalizations from the particular evidence reflected in the portfolio to the conception of competent teaching reflected in the standards on which the assessment is based. Or, more practically, “The key question is, ‘How likely is it that this finding would be reversed or substantially altered if a second, independent assessment of the same kind were made?’” (Cronbach, Linn, Brennan, and Haertel, 1997, p. 1). In addressing this question, we draw on two

kinds of evidence that are rarely available: comparisons of two different portfolios completed by the same teacher in the same year and comparisons between a portfolio and a multi-day case study (observation and interview completed shortly after portfolio submission) intended to parallel the evidence called for in the portfolio assessment. The case studies lasted 3 - 5 days, depending on each teacher's schedule. Consistent with Cronbach's (1988, 1989) "strong" program of validity, this study is explicitly *disconfirmatory*; it is intended to illuminate potential problems with assumptions about generalizability. Our formative goal is to raise issues that assessment developers and users can take into account in designing assessment systems and appropriately limiting score interpretations.

Conceptions of Generalizability

Messick (1989, 1996) characterized generalizability as "an aspect of construct validity" that is meant to "ensure that the score interpretation not be limited to the sample of assessed tasks but be generalizable to the construct domain more broadly" (1996, p. 250; see also 1989). He noted that generalizability has two important senses: (a) "generalizability as reliability … refers to the consistency of performance across the tasks, occasions, and raters of a particular assessment which might be quite limited in scope" (p. 250) and (b) "generalizability as transfer …refers to the range of tasks that performance on the assessed tasks is predictive of" (1996a, p. 250). Thus, inferences about the broader domain (in our case, competent teaching performance as defined by a set of standards) from a particular sample of evidence (as contained in a portfolio) can be productively conceived of in at least two distinct steps: from the observed performance to the more limited scope of what we will call the assessment domain (reliability) and then from the assessment domain to the outcome or standards domain (transfer or extrapolation). This distinction between kinds or levels of generalization is drawn by others as well, albeit with somewhat different language (e.g., Brennan and Johnson, 1995; Haertel, 1985; Haertel and Lorie, in press; Kane, Crooks, and Cohen, 1999) (Note 2).

Within psychometrics, generalizability has typically been evaluated in terms of quantitative indicators of reliability or transfer. These concepts from psychometrics will be useful--even though this is a qualitative study--for helping us frame and learn from the results of our comparisons. The comparisons we offer will in turn, suggest the limitations of conventional theory for illuminating the complexity of variations involved in teaching practice and making well warranted decisions that accommodate that variation.

This first level of inference (reliability) involves generalization from a set of representative observations to a well specified assessment domain (or universe of generalization) consisting of similar observations (Kane et al., 1999; Brennan, 2001). We are not simply interested, for instance, in how an examinee performed on a particular set of tasks on a particular occasion; rather, we are interested in estimating how an examinee would perform on tasks/occasions *like these*. Further, we want some assurance that the score is not based on the idiosyncrasies of a particular judge but that similarly qualified judges would likely interpret the performance in the same way.

Reliability is appropriately conceptualized and investigated as a faceted concept that encompasses multiple sources of "error" or variations over which we want to generalize (differences in tasks, raters, occasions, and so on that are intended as samples from the same assessment domain). A set of scores can have multiple reliabilities and errors of measurement depending on which sources of variation are taken into account. The appropriate domain of generalization, including the sources of variation over which we want to generalize, depends on the decision to be made (Cronbach et al., 1997). For those sources of variation over which we want to generalize, empirical studies that examine these variations—across tasks, occasions, raters, etc.—are required to support the generalization. As Brennan (2001) argued, the notion of "replication" is central to an understanding of reliability. Generalizability theory (Cronbach, Gleser, Nanda, and Rajaratnam, 1972; Brennan, 1983; Shavelson and Webb, 1991) is, perhaps, the most commonly used theoretical model that enables the effects of various sources of error to be "disentangled" and estimated simultaneously, although other models, especially those based on Item Response Theory (IRT) (e.g., Engelhard, 1994, 2002; Myford and Mislevy, 1995; Wilson and Case, 1997) are becoming more widely used (see Mislevy, Wilson, Ercikan, and Chudowski, 2002, and NRC, 2001a, for a discussion of alternative models). (Note 3) With generalizability theory, reliability is idealized as a statistical generalization based on "random" samples from the assessment domain. Brennan (2001) acknowledged that the notion of random sampling is an "idealization that is not fully supported", but noted that "the central conceptual distinction is not so much between fixed and random in the literal sense of 'random,' as it is between fixed and 'not fixed'" (p. 302). Reliability estimates can be quite misleading if a facet that varies in the assessment domain (possible essay prompts, for instance) is not included in the estimated error of measurement. An unfortunate practice is reporting reliability estimates for performance assessments based on differences among readers but ignoring potential differences among tasks even though the intended generalization is to a broader domain of tasks like these. This can seriously overestimate the quality of the generalization to the intended assessment domain.

Turning to the second level of generalization (transfer or extrapolation), this involves generalization from the more limited and carefully specified assessment domain to a broader outcome domain, which includes the full range of performances about which we would like to generalize. As Kane and colleagues noted, most educational concepts are quite broad; rarely are we interested simply in how examinees perform on other (test) items like these. Using reading comprehension as an (often cited) example, the outcome domain of interest might include a wide range of types and genres of text (e.g., newspapers, magazines, novels, instructional manuals, technical reports, text books, friendly letters, business letters, signs, forms, lists, tests), read for a variety of purposes, in many different contexts, requiring various kinds and depths of background knowledge to understand, with readings represented in multiple ways (writing, conversation, mental images or concepts, drawing, marks on answer sheets, and the like).

This level of generalization clearly spills over the bounds of reliability into validity more generally and typically involves a more tenuously warranted set of

inferences. Warrants for transfer generalizations include logical or theoretical arguments about the relationship between the assessment domain and outcome domain. A common approach “is to argue that the skills needed for good performances in the universe of generalization (e.g., problem definition, problem solving) are essentially the same as, or are a critical subset of, those needed in the full target domain” and “that anyone who performs well on the assessment should also be able to perform well in the target domain and anyone who performs poorly on the assessment should also perform poorly in the target domain…” or at least that “the skills being assessed are necessary (if not sufficient) for effective performance in the target domain” (Kane et al., 1999, p. 11).

Empirical studies supporting transfer generalizations might involve “criterion studies,” examining of the relationship between test performance and some “especially thorough (and representative)” sample from the outcome domain (Kane et al., 1999, p. 10) or, more practically, a “series of small experiments regressing various outcomes on test performance” (Haertel, 1985, p. 35). Given the near infinite range of possible studies, some means of deciding which are most important to undertake given limited resources is necessary. As Kane and colleagues noted, “in practice, the argument for extrapolation is likely to be a negative argument.”

> A serious effort is made to identify differences between the universe of generalization and the target domain that would be *likely to invalidate* the extrapolation. If no major differences are found, the extrapolation is likely to be accepted. If the impact of some differences on the plausibility of extrapolation is unclear, it may be necessary to check on their importance empirically. (Kane et al., 1999, p. 11)

Empirical Evidence of Generalizability

With Performance Assessments, In General

With performance assessments, the most commonly examined sources of error are those due to raters and tasks. Empirical studies of reliability or generalizability with performance assessments are quite consistent in their conclusions that (a) reader reliability, defined as consistency of evaluation across readers on a given task, can reach acceptable levels when carefully trained readers evaluate responses to one task at a time, and (b) adequate task or "score" reliability, defined as consistency in performances across tasks intended to address the same capabilities, is far more difficult to achieve (e.g., Breland et al., 1987; Brennan and Johnson, 1995; Dunbar, Koretz, and Hoover, 1991; Gao and Colton, 1997; Gao, Shavelson, and Baxter, 1994; Lane, Liu, Ankemann, and Stone, 1996; Linn and Burton, 1994; McBee and Barnes, 1998; Swanson, Norman, and Linn, 1995). In the case of portfolios, where the tasks may vary substantially from student to student and where multiple tasks may be evaluated simultaneously, inter-reader reliability may drop below acceptable levels for consequential decisions about individuals or programs (e.g., Koretz, McCaffrey, Klein, Bell, and Stecher, 1992; Nystrand, Cohen, and Martinez, 1993). Adequate levels of score (reader and task) reliability have typically been

achieved by further standardizing the task directions, choosing tasks with higher intercorrelations, disaggregating the portfolio into separate tasks that can be scored one at a time, and then estimating generalizability as one would with any collection of performance tasks. Brennan (2001) cautioned that tasks and raters are only some of the sources of error that are likely to matter. He cited other sources of variation that should likely be taken into account. These included different occasions, both occasions of testing as well as occasions of scoring, and different methods of testing as sources of error. He noted that some of these, such as different methods, are better conceptualized as convergent validity studies (rather than as reliability studies per se). (Note 4) Of course, certain types of estimate are often deemed not feasible, including parallel forms reliabilities with portfolios and assessments of performance in different contexts (ETS, 1998; Harris, 1997; NRC, 2001b; Porter et al., 2003).

Special studies involving performance assessments have looked at relationships among methods of assessment: between multiple choice and performance assessment (e.g., Lane et al., 1996; Crehan, 2001); (Note 5) between different methods of performance assessment, such as direct observation of scientific experiments and analysis of students notebooks (Shavelson et al., 1991); and between on-demand and school based tasks (Gentile, 1992, in Brennan and Johnson, 1995). The general conclusion is that different methods appear to be getting at somewhat different constructs (e.g., Brennan, 2001; Brennan and Johnson, 1995). Fewer operational assessments in education undertake this sort of empirical research, relying instead on empirical evidence of reliability and logical arguments about content-relevance and representativeness. And, indeed, while the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999), require at least some sort of empirical evidence about reliability, they mention “external validity” as only one potential source of evidence, but leave the choice of validity evidence up to the assessment developer and user.

Some authors note a tradeoff between these two levels of generalization. Strengthening the faithfulness with which the assessment represents the outcome domain often undermines the reliability of assessment (as reflected in the many technical problems with performance assessment) and enhancing reliability, for example by employing a larger number of shorter tasks, undermines fidelity (e.g., Kane et al., 1999).

With Teaching Performances, in Particular

Research into the generalizability of performance assessment of teaching has tended to emphasize much the same sort of evidence described above, focusing primarily on consistency among tasks and judges. There are two major programs of research that are most relevant to our study, the portfolio assessments of the National Board for Professional Teaching Standards and the observation/interview assessments of Praxis III. Both of these assessments are developed by the Educational Testing Service.

National Board’s standards-based assessments are designed to certify the accomplishment of experienced teachers with at least three years of service. Assessments are developed or underway for over thirty different certificates

(differentiated by subject area and age of students taught). The ten performance tasks that comprised the assessment in each certificate area (when the research described here was undertaken) are divided into two parts: a portfolio completed by candidates in their home schools across a year and a one-day assessment-center experience. The school based portfolio consists of (a) four tasks that ask candidates to document their practice, through videotapes and samples of student work, and to provide "extensive analytical and reflective commentary" (Pearlman, in Jaeger, 1998, p. 191), and (b) two tasks that ask candidates to document their accomplishments outside the classroom and explain why they are important. The four assessment-center tasks provide candidates with materials such as student work samples, assessment records, instructional resource materials, or professional reading and ask them to use the materials to diagnose the status of student learning, plan instruction, and so on (Pearlman, in Jaeger, 1998, p. 191). (Note 6) Each exercise is scored independently by two reviewers. The resulting scores for each exercise are weighted and aggregated to form an overall composite score for each candidate. This composite is then compared to a predetermined passing score.

The National Board's *Technical Analysis Report* (ETS, 1998) described four relevant sources of error:

- Assessors: Would a candidate, given a different set of assessors, fare similarly on the assessment?
- Exercise Sampling: Would candidates perform similarly on a different set (sample) of exercises?
- Assessment occasions: Would candidates fare similarly if they took the same assessment on a different occasion?
- School context: Would candidates fare similarly if they happened to teach in a different school?

They noted that it is not feasible for them to provide evidence of reliability across school contexts or assessment occasions. With assessment occasions, they argued that there is likely to be a learning effect such that one would expect a candidate to fare differently (better) and so reliability may not feasibly be assessed.

They provided empirical evidence with respect to assessors and exercise sampling— concluding that both are adequate to support the assessment for its intended use (ETS, 1998, p. 125; see also Myford and Engelhard, 2001). (Note 7) With respect to exercise sampling, they cautioned readers about the limitation of such evidence since the set of tasks was explicitly designed to represent a multidimensional domain:

> Whether an assessment with the current design can be considered to allow for alternative forms in a traditional measurement sense is debatable. It is possible to argue that the exercises are but one possible sample from a larger domain of accomplished teaching or that the exercises, for all intents and purposes, comprise a fixed assessment of accomplished teaching. (ETS, 1998, pp. 107-108)

This is, in fact, typical of the way in which task generalizability is investigated

with portfolio assessments (e.g., Klein et al., 1995; Koretz et al., 1992; Reckase, 1995; Nystrand et al., 1993); what we have is an estimate of internal consistency (based on tasks that were designed to access quite different elements of teaching practice) and that treats as fixed a wide range of factors that may in fact vary. Following Brennan (2001), this is not really a replication "using two full length operational forms" (p. 313).

With respect to "transfer," Bond, Smith, Baker, and Hattie (2000) examined the relationship between scores on the National Board's assessment (in two certificate areas for 65 teachers) and 1-3 hour observations of teaching accompanied by interviews with teachers and some students. The casebooks produced from the visits were scored according to thirteen dimensions of accomplished teaching identified in an extensive literature search. Using discriminant analysis, they were able to correctly classify 84% of teachers as to whether they had been certified using the National Board's assessment. Other studies are currently underway (see www.nbpts.org). While the National Board's goal was primarily documenting consistency across the sources in support of the validity of the NBPTS assessment, our purpose is to illuminate both similarities and differences at the level of particularity that qualitative methods allow.

ETS's PRAXIS series, which is intended for use with beginning teachers, involves three sets of assessments: PRAXIS I focuses on basic skills, PRAXIS II on content knowledge and general pedagogical knowledge, and PRAXIS III on teaching performance. The PRAXIS III assessment involves direct observations of classroom performance over a series of "assessment cycles." An assessment cycle consists of a preliminary description of the context, the students, and the lesson-to-be-observed, prepared by the beginning teacher; an observation of a lesson of instruction by a trained assessor (experienced teacher); and pre and post semi-structured interviews. The assessor's notes are then scored on a list of nineteen criteria (that were developed through an extensive literature review and job analysis survey) and an overall score given. "Summative decisions are made based on cumulated data from two or more assessors based on two or more assessment cycles" (Dwyer, 1998, p. 8). In addition to the obvious differences in methods, PRAXIS III is intended for use across grade levels and subject areas, and the criteria for classroom observation have not been tailored to particular subject areas as with INTASC and the National Board. Although this leads to a somewhat different emphasis, Porter et al. noted the similarity of the PRAXIS criteria to the general principles of the National Board and INTASC. While there are multiple studies of assessor reliability, there are no reports of generalizability across assessment occasions that we could locate (Dwyer, 1998; Myford and Lehman, 1993; NRC, 2001b; Porter, Youngs, and Odden, 2003; Myford, personal communication, 3/5/03; Wylie, personal communication, 5/2/03). With respect to generalizability across occasions, assessment developers caution:

> "The purpose and consequences of the assessment, particular local circumstances, and the beginning teacher's level of performance (both absolute and in terms of improvement) are factors that determine how many assessment cycles will be carried out. Guidelines governing Praxis validity and use prohibit

> decision-making on the basis of a single assessment cycle or on the judgment of a single assessor (Educational Testing Service, 1993b).” (Dwyer, 1998, p. 171)

Thus, the comparisons in the study reported here--which involve full length replications of portfolio assessments, methods of performance assessment, and classroom contexts in which the same tasks can be implemented--begin to address an important gap in our understanding of the generalizability of portfolio assessments of teaching and, perhaps, of performance assessments more generally.

Research Design

Our study draws on qualitative methods to address questions of portfolio generalizability through comparative content analyses across different portfolios and different methods of assessment for the same teachers. Consistent with Kane and colleagues’ (1999) conception of a negative argument, built from a serious effort to disconfirm, our goal is to illuminate differences that challenge assumptions about generalizability. Where to locate these comparisons in terms of the level of generalizability described in the previous section is an open question. At face value, one might argue that the portfolio-portfolio comparison is a reliability issue (different occasions on which same tasks are performed), and the portfolio-case comparison is a transfer issue (different methods and different occasions). And yet, as we return to this issue after sharing our findings, the nature of variations that the different occasions afford makes this problem far more complex--as occasion is confounded with uncontrollable aspects of context--and raises important questions about the nature of the assessment domain to which we can appropriately generalize. These are the variations that can be invisible when portfolio reliability is examined via intercorrelations among tasks and readers.

We begin with a brief description of the INTASC portfolio assessment system and then describe data collection for the two comparative studies--portfolio-portfolio comparison and case-portfolio comparison--which were replicated in secondary English Language Arts (ELA) and Mathematics (Math). Since the comparative content analyses for both studies follow a similar pattern, we describe those activities in a fourth section. While the data sets are small from a quantitative perspective (29 comparative cases across the two studies and two subject areas), our goal was to understand each comparative case in depth and to illuminate issues for assessment developers and policy makers to consider.

INTASC Portfolio Assessment

The portfolio assessments are intended for teachers in their first, second, or third year of teaching. To guide the portfolio assessment, INTASC has developed a set of general and subject specific standards based on INTASC's Principles for Beginning Teachers and standards from the relevant professional communities. The standards and related assessments are intended to provide a coherent developmental trajectory with those of the National Board. The assessments ask candidates for licensure to prepare a portfolio documenting

their teaching practice with entries that include: a description of the contexts in which they work, goals for student learning with plans for achieving those goals, lesson plans, video tapes of actual lessons, assessment activities with samples of evaluated student work, and critical analysis and reflection on their teaching practices. Unlike the National Board portfolios (which contain four separate entries), these entries are organized around one or two units (8 – 10 hours) of instruction such that the portfolio cannot easily be broken into parts for separate evaluation. Judges evaluate the portfolios in terms of a series of "guiding questions" focused on the portfolio but based on the standards described above; they record evidence relevant to each guiding question and develop interpretive summaries or "pattern statements" that respond to the question; then they determine an overall decision about the candidate (Note 8). As developed by INTASC, the portfolios were intended both for professional development and for informing decisions about licensure. Of the 10 INTASC states that participated in the development of the portfolio assessment, only Connecticut is currently using it to inform licensure decisions.

For this study, participants were recruited **from fieldtests in multiple INTASC states** in 1998-2000. Because our interest in this paper is about the generalizability of portfolios for licensure decisions, we chose to evaluate the portfolios using the guiding questions and decision guide as they were used by Connecticut for field tests in 1999-2000, even though the participating teachers were recruited from multiple states. As it was implemented in Connecticut in 2000, there were four possible levels to the overall decision: conditional, basic, proficient, and advanced. Judges also completed a "feedback rubric" on which they selected performance levels that best characterized the portfolio with respect to each guiding question. The assessment occurred as part of a 2-3 year induction program in which beginning teachers who had an initial three-year license were provided with a mentor in the first year and the opportunity to attend state-sponsored workshops to prepare them for the assessment. When fully operational, teachers who did not pass the portfolio assessment in their second year would continue in the program for another year. If they did not pass in the third year, they would be required to reapply for the initial license after successfully completing additional course work or a state approved field placement.

Portfolio-Portfolio Comparison Data Collection

A small sample of secondary beginning teachers in math (n=7) and ELA (n=6) were recruited to complete two portfolios during the same year, choosing classes and units of instruction that differed as much as possible within their routine teaching assignments. They were compensated for the second portfolio. Not surprisingly, it was very hard to find beginning teachers willing to assume the burden of two portfolios, and it is impossible to fully understand how these stalwart volunteers might have differed from their colleagues. We can say that their portfolios do reflect a range of performance levels, teaching practices, and school contexts and that their paired portfolios do illuminate an instructive array of differences, consistent with the goals of the study.

Case-Portfolio Comparison Data Collection:

Another small sample of secondary teachers in math (n=8) and ELA (n=8) was asked to allow case studies of their teaching shortly after they submitted their portfolios. The sample was recruited to include differences in gender, ethnicity, school context, and performance level (based upon a quick read through by the portfolio developers). The case studies took place over 3-5 days (depending on the teacher's schedule) during which researchers observed classes; conducted entry, exit, and brief daily interviews with the teacher; and interviewed the school principal and, if possible a mentor, regarding the support available to the teacher. [See Ball, Gere, and Moss, 1998; Moss, Rex, and Geist, 2000a, 2000b for fieldwork and case write-up guidelines.] Case study researchers observed two classes: the class used in the preparation of the portfolio and a second class. As with the portfolio/portfolio comparison, we asked for a class that differed as much as possible within the teachers' routine teaching assignment (but sometimes we were only able to observe a different section of the same class). Our intent was to parallel the information collected in the portfolio as closely as possible and to gather additional information about the teacher's background, school context, and experience preparing the portfolio to address additional questions of fairness. Teachers were given a small honorarium for participating in the case study. As before, it is not possible to know how these volunteers differed from the larger population of beginning teachers.

Case study researchers, all experienced teachers in the appropriate field, were taken through an abbreviated course of study (with practice and feedback) in taking fieldnotes and conducting interviews relevant to the project. Tape recordings and artifacts were used as back-up. Field and interview notes were read by a senior researcher and questions of clarification and elaboration were raised to guide revisions (which could be supported with audio-recordings and artifacts). Case study researchers were then asked to draw on their notes in responding to the Guiding Questions used to evaluate the portfolios. Again, a senior researcher reviewed the responses (with fieldnotes at hand) and raised questions to facilitate revision.

Comparative Analyses

The comparative content analyses for both studies were undertaken in a similar fashion. Research assistants (experienced teachers in the content area with graduate research training) used the guiding questions (and the dimensions contained in the related feedback rubric) to develop a coding scheme for the two sources of evidence. Videotapes were roughly scripted for coding. Then, answers to each of the guiding questions were developed for each source based upon a comprehensive review of the evidence, including the search for counter examples to challenge developing interpretations. Similarities and differences were then noted, organized by guiding question and overall. Justifications for perceived differences in performance level with respect to the criteria were developed. For the portfolio-portfolio comparisons in ELA, each pair of portfolios was read twice, in reverse order, by two research assistants, who then met to develop a consensus on any differences. (Note 9) For the portfolio-case comparisons and the portfolio-portfolio comparisons in math, a single comparison document was developed, and the process was audited by another researcher. The comparative content analyses typically took 3-5 days

per teacher and generated 30-70 pages of text each. These comparisons were then condensed into 2-3 pages versions that highlighted substantial differences both at the level of the guiding question and overall.

It is important to note that we have, for the purposes of this paper, bracketed questions about consistency among readers. Elsewhere we address concerns about differences in the way knowledgeable readers evaluate portfolios in different social settings when trained to reach consistent decisions and when allowed to draw on their own criteria of competent teaching (Moss and Schutz, 2001; Moss, Schutz, Haniford, and Miller, in preparation; Schutz and Moss, in press). Here, we present findings whose validity is based upon in-depth analyses, in which relevant differences in perspective between readers were resolved through consensus seeking dialogue. The issue for us is not the validity of a specific score; rather it is the validity of an interpretation of difference between two portraits of teaching and an argument for whether the observed differences are likely to matter in light of the evaluation criteria. We present our evidence for which differences are likely to matter in sufficient detail that readers can reconsider these judgments for themselves.

Structural Differences Between Data Sources and Asymmetrical Questions of Comparison

By structural differences, we mean those differences between data sources that could be anticipated in light of the different methods and which are, in fact, typically present in our data. With respect to the portfolio-case comparisons, beyond the obvious differences in data collection methods, it is important to note the following. While we attempted to have case study researchers present on days when teaching consistent with what is expected in the portfolio was occurring, it was not always possible to observe all the aspects of teaching called for in the portfolio. For instance, while the ELA portfolio required evidence of students' response to literature and students' processes in writing, the lessons observed in the case study might not cover both areas. The case-based evidence is typically weak with respect to formal assessment procedures since often no formal assessment was occurring. However, the case study provides substantially more evidence about daily classroom interactions. The case also provides rich information about the context in which the teacher worked and about teaching practices not foregrounded in the portfolio evidence. With respect to the portfolio/portfolio comparisons, the portfolio completed second is invariably shorter, often considerably so. It contains typically fewer artifacts and shorter commentary (sometimes with reference back to the first portfolio). This caused us to develop *an asymmetrical comparison and research question*:

> To what extent does the second portrait (case study or second portfolio) cause us to reconsider the evaluation of the teacher's performance in (what we'll call) the primary portfolio?

Findings

Our comparative analyses were set up to uncover differences in the two portraits of beginning teachers and to evaluate whether the differences were

likely to result in different decisions in light of the INTASC standards (as instantiated in the guiding questions and the decision guide as adapted and used in Connecticut in 2000). We make no attempt to estimate the frequency with which these sorts of differences are occurring; our evidence is not appropriate for that purpose. Again, our formative goal is to illuminate issues for assessment developers and users to consider in designing an assessment system, characterizing the appropriate domains of inference, and limiting interpretations appropriately.

We present our findings in the following sections: We begin with an overview of the variations in context of the classes and units selected by these teachers. Then, we illustrate our comparison methodology in substantial detail with comparisons in both math and ELA. In the first comparison, we provide an example of a case in which the differences observed do not seem to matter in terms of the relevant criteria (which was, we should note, true in the majority of cases). In the second comparison, we provide an example in which the evidence in the portfolio is, we'll argue, ambiguous, because the artifacts (videotapes, handouts) only partially support the written representation of the class; the case appears to clarify the ambiguity. Whether this is a difference that would "matter" depends on how the portfolio readers weigh the partially conflicting evidence. Thus this comparison, even more so than the first, illustrates some of the interpretive problems we encounter with these sorts of data— problems that we have tried to address through far more in-depth readings than would be possible in operational use. Then, we present a series of briefer vignettes that describe situations in which the second portrait caused us to question the conclusions we drew from the primary portrait.

Contextual Variations

As indicated above, for both sets of comparisons, we asked teachers to choose a second class that differed as much as possible within their routine teaching assignments. The classes they selected are presented in Tables 1 and 2 for secondary ELA and math teachers respectively. Given their selections, it is important to note the many different kinds of (often intersecting) contextual variations that are present in the comparisons we examine. These include: different sections of the same course (which entail differences in time of day and whether the teacher has taught the lesson before); differences in (perceived) ability levels and groupings of students, including those designated by the school directly (remedial, AP, and the like) or indirectly (scheduling in ELA resulting from math assignments) and those perceived by the teacher; different courses; different grade levels; different units within the same course; differences in (mix of) cultural backgrounds of students; different times of year (which involves differences in teachers knowledge of and relationship to students); differences in class sizes; differences in availability of curriculum and support materials; differences in extent to which these materials are consistent with the standards. These are all variations that are fixed for a given portfolio assessment of a teacher and are unexamined when all we have is the single set of performances from a given class. (Note 10)

Illustration of Analysis in ELA with a "Complementary" Portfolio/Case Comparison

To illustrate our comparison methodology in ELA, we focus on one portfolio/case comparison, "Ms. Bertram (Note 11)," in which the activities we observed differed substantially, and yet we found the portrait of the teacher conveyed in the case study provided quite consistent evidence with respect to the general evaluation criteria. We illustrate this comparison in some detail both to document our practices of analysis, and to show how two quite different activity contexts can nevertheless support similar conclusions about the teacher. We begin with a discussion of the ELA portfolio guidelines, the guiding questions (developed by INTASC and revised by Connecticut) to evaluate completed portfolios, and the way in which we applied them for this study. Then we return to the specific case of Ms. Bertram.

The ELA portfolio handbook asks candidates to complete two distinct entries: one each in teaching response to literature (RL) and processes of writing (PW). Teachers may choose the same class or different classes for these two components. Across these two exhibits, we have the following sources of evidence from each teacher: (a) the teacher's rationale for her choice of literature and writing assignment, (b) the teacher's daily logs for 10 lessons in which she describes the activities she and the students engaged in (providing copies of instructional artifacts) and writes brief reflections about how the day's lesson went, (c) video tapes of two-three activities reflecting different participation structures (d) teacher's reflections on the videos, (e) five samples of student writing, including multiple drafts, with teacher's comments on the writing, (f) the teacher's reflections on the students' writing, and (g) the teacher's general reflections on her teaching in the unit.

In the case study, we have fieldnotes depicting the activities and the discourse in the classroom across three days for each of the two classes. Through notes from a series of interviews, we learn about the teacher's goals, her specific plans for daily lessons, her reflections on how the lessons went (in general and for particular students), and her goals for professional development.

The guiding questions for ELA (initially prepared by INTASC and revised by Connecticut for use in 1999-2000) are organized into four separate categories. (1) Questions about *literacy* focus on "connections among responding, interpreting, and composing" with an emphasis on the extent to which students develop their own meanings. (2) Questions about *instruction* focus on how the teacher organizes students' learning--including questions about alignment between goals and instructional strategies, about integration of activities within and across lessons, and about materials--with an emphasis on the extent to which instruction provides learning opportunities (challenges) for all students and promotes independence. (3) Questions about *analysis of learning* focus on formal and informal assessment of students' work--how the teacher monitors students' progress, communicates with them about their learning, and uses that information to inform instruction. (4) Finally, questions about *analysis of teaching* focus on how the teacher reflects on student learning and uses that reflection to inform her practice. (Note 12)

While some of the guiding questions were quite specific and descriptive in nature (e.g., *"Describe how the teacher helps students use a writing process, including context, purposes, and conventions of standard written English."*);

others involved much higher levels of inference that required integrating multiple types of evidence (e.g., *Describe the ways in which the teacher creates a learning environment that provides all students with opportunities to develop as readers, writers, and thinkers.*"). In analyzing the portfolios, we found ourselves following a multi-step process. We began with describing the various sources of evidence (e.g., describing the teacher's goals, outlining the progression of lessons, scripting the videotape, characterizing the artifacts in terms of the nature of students' responses and any written comments by the teacher; illustrating the ways in which teachers reflected on their students' work in their commentary). Then we developed interpretations that coordinated various sources of evidence (e.g., considering the relationship between the teacher's goals and the progression of lessons to evaluate alignment and scaffolding or between the teacher's commentary on the video and what we had observed to evaluate quality of reflection). Finally, we moved to the level of responding to some of the higher-inference guiding questions (e.g. "*Describe how the teacher uses knowledge about students to meet their needs in instruction and provide them with opportunities to learn*" or "*Describe the ways in which the teacher creates a learning environment that provides all students with opportunities to develop as readers, writers, and thinkers.*"). For the case studies, the fieldnotes from the classroom observation and notes from a series of interviews with the teacher allowed us to engage in much the same process. Our task was somewhat easier as the case study writer had constructed responses to the guiding questions that drew on evidence form the field and interview notes. We nevertheless reviewed the field and interview notes in light of the case study writer's conclusions and often included the additional detail in our comparisons.

The Appendix provides brief excerpts from the 70-plus page portfolio/case comparison document prepared by LeeAnn Sutherland. It shows brief examples of the sort of evidence we have from the two methods and illustrates the way we have combined the evidence to develop interpretations and comparisons relevant to the guiding questions. Below we offer some general conclusions based both on the comprehensive evidence in the longer document for which the Appendix provides only brief examples.

Ms. Bertram teaches sixth grade English Language Arts in a middle school located in what the case study writer describes as a small, relatively affluent suburban community. In the portfolio and the case, we see a "reading and writing" class of 24 students who meet daily for two periods. In addition to this reading and writing class, the case study writer observed another section that covers only writing.

For the response to literature exhibit in her portfolio, Ms. Bertram selected a series of lessons based on the study of a novella commonly used with this age group. Three separate tasks require students to (a) identify the character traits of the main characters, (b) compose a written response citing which character they felt they were most similar to/could relate to best/liked the best and why, and (c) use that reflection as the foundation for creating a simulated journal. In the processes of writing exhibit, we see Ms. Bertram guide students through the development of a poem using metaphors to describe their mothers in preparation for Mother's day.

The case study describes three days of parallel lessons in two classes where the teacher focuses on having students select three best pieces of writing from their notebooks, complete evaluation sheets about each one, exchange with a partner who would name his or her choice for the writer's best piece on a 'Nomination Ballot,' revise that piece of writing, and publish it on a web page.

Even though the activities are substantially different, there is nothing in the case that would cause us to question our evaluation of the primary portfolio. Both portraits show the teacher using a variety of activities to help students use literature to make connections, take others' perspectives, and explore concepts, scaffolding their learning through the activities she creates and the discussion she guides. She also uses a variety of activity structures (e.g. small group, whole class). We have ample evidence of similar classroom interaction wherein the teacher poses questions to which students respond initially (to begin an activity or class session), and she then builds from students' responses to guide subsequent questions, consistently validating their contributions. In both portraits, we see the teacher employ a variety of strategies to guide students in developing as readers, writers, and thinkers. Either portrait would tell us that this is a highly reflective teacher who uses that reflection to shape practice immediately and to think about changes in her future practice. She consistently addresses both the strengths and weaknesses of each lesson as well as their relationship to the larger unit. Thus the evidence in the case reinforces the conclusions from the portfolio in somewhat different contexts of teaching.

Illustration of Analysis in Math with an Ambiguous Portfolio and Clarifying Case

Complex evidence of the sort contained in the portfolio and case often presents substantial interpretive problems to readers. While this is not the focus of this paper, reading problems do impact the nature of the conclusions we draw. Here we illustrate our analytic practices with a math comparison and present a situation where the evidence in the portfolio is somewhat ambiguous and where the additional evidence in the case appears to support one potential portfolio interpretation over the other.

The math portfolio handbook focuses on a single 8 – 12 hour unit in mathematics and requests similar artifacts as requested in the ELA handbook. Here the portfolio contains a description of the classroom context, descriptions of a series of lessons with instructional artifacts (e.g., handouts, assignments); videotapes, student work, and reflections on two featured lessons; a cumulative evaluation of student learning with accompanying reflection; a focus on three students across the featured lessons and cumulative evaluation of learning; and analysis of teaching and personal growth. As with the ELA handbook, then, we have partially independent artifacts (including the videotape, instructional artifacts, and samples of students' work) against which to evaluate (some parts of) the teacher's description and reflection/evaluation on what happened.

The guiding questions in mathematics are organized into five categories (as initially prepared by INTASC and revised by Connecticut for use in 1999-2000). (1) *Tasks* focuses on the appropriateness (variety, richness, challenge,

accessibility) of the tasks selected by the teacher and on how effectively they are implemented (clarity, accuracy, alignment, and responsiveness to students' interests, styles and experiences). (2) *Discourse* focuses on how effectively the teacher orchestrates discourse, uses tools and materials to support discourse, and promotes discourse among students in which powerful kinds of thinking predominate (defined as students exploring a variety of approaches to problems and explaining their reasoning with evidence). (3) *Learning environment* focuses on how effectively the teacher manages the physical, time, and social aspects of the classroom and encourages participation and engagement by all students. (4) *Analysis of learning* focuses on how effectively the teacher assesses students' learning (accuracy, variety, and alignment with objectives and tasks) and communicates with students about expectations and feedback. (5) Finally, *analysis of teaching* focuses on how the teacher learns from and improves teaching. The comparison methodology was similar to that presented in ELA. (Note 13)

The mathematics teacher in this portfolio/case comparison works at a large urban high school, in which 68% of students receive free/reduced lunch. The portfolio presents an 11th grade Integrated Geometry course. The teacher, Ms. Fleming, explains that this course is the lowest level geometry course offered by the school and that she closely follows the text. The unit presented in the portfolio concerns tessellations and triangles. The case study follows the same Integrated Geometry class and a 9th – 10th grade Math I class. Ms. Fleming reports that Math I is the lowest level math course offered by this school with the exception of remedial math. At the beginning of the course the students shared textbooks with another class; however, Ms. Fleming indicates that these texts soon disappeared. Ms. Fleming uses worksheets left by a previous teacher and generates her own curriculum worksheets. She reports that approximately 75% of Math I students are failing. The lessons presented in the Math I class focus on basic arithmetic, naming of geometric objects and measurements. The Integrated Geometry lessons observed for the case focus on triangles, angles, and parallel lines. The case was conducted late in the year when both classes were reviewing material for a final exam.

We begin with an extended discussion of the portfolio because it alone raises a complex interpretive problem when the partially independent artifacts (videotapes, handouts) are compared to the teacher's descriptions of what is happening. In the interest of space, we focus on connections across lessons, nature of mathematical tasks, and the implementation of tasks in classroom discourse. Then we turn to the case where the portrait of the teacher is substantially different from what is portrayed in the written portion of the portfolio. [Both descriptions draw heavily on Pamela Geist's extended comparison document.]

The evidence in the portfolio creates a picture of a teacher that sees how the mathematics of a unit connects across ideas and to prior and later learning. Early in the portfolio, Ms. Fleming describes some of the mathematical connections she believes are important for students to understand. She writes,

> *Knowing the properties of triangles is important to the student of mathematics because it is the starting point for learning the*

properties for special triangles and enclosed figures, namely polygons. For example, the Triangle Angle-sum Theorem can be used to derive the sum of the interior angles of a quadrilateral and convex polygons. It also lays the foundation for students to learn about pyramids and other three dimensional figures.

Ms. Fleming describes in detail how the seven lessons across the unit connect mathematically and what students will learn across the unit to accomplish learning goals and objectives. For example, she explains,

It was important to show the relationship between the exterior angle of a triangle with the adjacent interior and remote interior angles. Once the properties of a single triangle were established, it was necessary to establish the relationship between a pair of congruent triangles and how to use the postulates to establish congruence. In order to establish this, students had to learn how to make congruence correspondence and congruence statement. Of course this also leads us to establishing proof, but my department recommended not to introduce proofs with this level class.

In the portfolio, Ms. Fleming develops a strong case for the predominance of discovery-type tasks and learning. She explains that hands-on discovery type tasks dominate her practice and that students learn best in these types of lessons. She writes,

The tasks that are most effective are of a 'discovery' or 'hands on' type… [explaining] *when my students "see it" and "find it" the learning is retained …. I try to let my students have the experience of discovery even when it seems small. I have found that using the discovery method works best for my students so I have tried to use it often…*

Using this method students get to see ideas. For example, when students put together the angles from a triangle and actually saw that it made a straight line, they knew that adding the interior angles of a triangle would equal 180°. They saw that it worked for all triangles regardless of the size and shape.

In the portfolio artifacts, we see a range of tasks including tasks that appear to offer opportunities for discovery and those that focus more on recall and application of definitions and facts. Consistent with the teacher's description, a series of problems presented in one of the instructional artifacts asks students to work at drawing and measuring the various angles within the triangles and record their data. The questions that follow ask students to detect a pattern in the data and develop statements or conclusions about various relationships within the triangles. There is much writing in the portfolio explaining that these tasks and others like them are selected because they support students' opportunity to formulate conjectures, reason about mathematical ideas, and justify results. The teacher also provides evidence of other types of tasks that are designed to check on students' general understanding of geometric shapes and their properties. For instance, they ask students to classify shapes, recall definitions and theorems, use definitions and properties to find other measures,

and to justify answers with a known theorem or definition. The tasks appear on daily activity worksheets, homework assignments, and on assessments such as tests or quizzes.

We turn next to the videotape to see how these sorts of tasks are implemented in classroom interaction. Here, we focus in detail on one of two videotaped lessons providing excerpts from our rough transcript of the videotape and Ms. Fleming's reflections on what occurred. The videotape is less effective in making the teacher's case for discovery-type learning. We see Ms. Fleming guiding the discourse with students responding in short statements that restate a definition or fact. The excerpt we've selected begins about 4 minutes into the tape after Ms. Fleming has finished reviewing, through brief question and answer segments, the previous day's lesson for students.

T Okay. So let's look at a triangle (she has an example on the overhead - (4:37 into the tape)). We have remote interior angles, we have exterior. We have adjacent interior. Let's look in relationship to this one angle (points to image on the screen). Here we have an exterior angle. It's outside the triangle. The adjacent interior is the one that is what?

S Sharing the same sides.

T Sharing the same sides. So it's adjacent. Adjacent means?

S Next to.

T Next to, okay? Remote means, we said?

S Far away.

T So these two are far away from this exterior angle. Right? These are going to be your remote interior.

S (unintelligible)

T Now look at this, if I said 2 is you exterior angle, what is the adjacent angle for 2. Where would it be located?

S (A student is asked to come up and point out the specified angle. Other students were calling out some helpful comments as well as "I know, I know" as he points to various angles the teacher asks him to identify - remote interior. She asks him to confirm that he is pointing to remote interior or remote exterior. He confirms remote interior).

T Very good. So depending on which angle you pick, your remote interior angles will be switching sides. Okay, this is my exterior angle 1, right (she points)? So what angle is adjacent to that and inside the triangle?

S What's adjacent?

T Adjacent means next to, it's touching. It's sharing the same side. So angle A over here would be CAB. Correct? CAB is adjacent to angle ...? I'm looking

at angle 1. What is it?

T I'm looking at angle 1. What kind of angle is angle 1?

S Exterior

T What is the adjacent angle to angle 1?

S 4

T What are the remote interior angles?

S 5 and 6

T 5 and 6. Much better.

T & S (At 7:28 in the tape) (Some students need clarification so students and teacher have a brief discussion on the different types of angles and their relationships to each other).

T (Teacher moves around the triangle she has on the overhead and asks for students to quickly identify remote, adjacent, and exterior angles).

T Now, today you're going to look at the relationships between these angles. Okay? I'm going to hand out a worksheet and you're going to do that.

[Break in sequence. Students are now working together in small groups on the assigned worksheet. The teacher walks around to answer questions and check their work. Students are comparing work. They use rulers and protractors to measure angles and help each other construct the various triangles on the worksheet.) (It's hard to hear what students are saying to one another but the teachers voice can be heard from time to time.)]

In reflecting on this lesson, Ms. Fleming describes how she interacted with students to arrive at a solution:

> *I did not offer 'answers' for the students, but guided them using questions to arrive at a solution. For example, when the male student attempted to identify the angles on the transparency, I realized that he was trying to 'bluff' his way out of it. I guided him by repeating the names of the angles, emphasizing the words adjacent and interior.*

And about the small group time, she writes:

> *When a student asked me if she measured an acute angle correctly, (she did not by reading the protractor incorrectly), I asked her if her angle was greater or less than 90°, and if her answer made sense. When student A asked me about the measures of her angles, I asked her how she could check them. Once she 'got it', she proceeded to help another member of her group.*

This is not an unreasonable representation of what occurred, and it helps us understand why she made some of the choices she did. Viewed in light of this interpretive commentary, and taken together with description of all the lessons that reflect a privileging of discovery-type learning, it is possible to situate the evidence in the videotape within a larger picture that mitigates its dominant impression. While not presented here, the other videotape and reflections surrounding it raise similar issues; the teacher's description surrounding the lesson creates a different image than what we might infer from the videotape alone.

As teachers, we know that even in the most 'learner-centered', discovery-oriented classroom, there are often (with good reason) stretches of dialogue that resemble what we see here. That we can't hear what is happening among the students on the videotape allows the teacher's characterization to shape our impressions. Viewed alone, this portfolio can be constructed as a relatively strong performance, better than just passing, even though the evidence provided by the artifacts is a bit uneven.

Turning to the case study, what we see reinforces what we see in the videotape and, taken together with what the case study researcher reports from his interviews with the teacher, presents a substantially different portrait. We focus on the same aspects of the teacher's practice, presented in essentially the same order: connections across lessons, nature of tasks, and implementation. About her characterization of connections across tasks, the case study researcher writes: "In the pre-lesson interviews when asked to describe her objectives for the next day's lesson, they were always in terms of discrete topics to be covered, sometimes by book chapter. For instance, Ms. Fleming explained her plans for her Math I class, "*I am presenting material that is very close to that they are seeing on the exam. I will do Chapter 10 tomorrow, reading graphs, finding mean, median, mode, and range*." Or, following a Geometry lesson, she says: *"they can do triangles, but not the parallel lines. I keep throwing these [parallel line problems] at them so they keep seeing them."* In his search for counter evidence about this developing pattern, the case study researcher offers the following quotation:

> *We do introduce the concept of showing how triangles can be congruent and we will ask them to give reasons. The last thing they were doing was perimeter and area for rectangles, parallelograms, and other quadrilaterals. We did Pythagorean Theorem and area under the curve using a trapezoid. And we try to reason with them by making 'cubes' under the curve and having them count the 'cubes'.*

The case study researcher argues that the teacher merely makes mention of ideas that were presented in earlier lessons or other contexts but did not offer any deeper understanding of how ideas are connected, only that they are.

The case study researcher develops a very different portrait of the dominant kinds of tasks offered and the kinds of learning they promote. The case study writer concludes, "there is little diversity or richness in the problems offered," and, "the majority of tasks are one-step applications of definitions and theorems." He describes, "Both in the integrated geometry and in the geometry

content of the Math I course, she emphasized fundamental skills such as naming and applying simple definitions. In the first period I observed of Integrated Geometry the first set of problems all are based on knowing definitions (e.g. altitude, median, congruent) or theorems (e.g. corresponding angles congruent). The questions are one-step applications of definitions where the only probe mentioned in the problem, 'How do you know?"?' is more a reference to naming the correct specific theorem used to solve the problem." He presents numerous examples of these kinds of tasks. He concludes, "Ms. Fleming's objectives across the tasks she offered were centered around coverage of facts, definitions and theorems students had memorized and not on the development of particular skills or understandings of broader concepts."

The case study writer offers a description of the typical discourse pattern, "There was one dominant pattern of interaction around the tasks offered in Ms. Fleming's classes. My characterization of this pattern is based on three observations in the context of two different mathematics classes– Integrated Geometry and Math I. Ms. Fleming offered students a set of problems, similar to those given earlier, in the form of a worksheet. Ms. Fleming engaged students in a Question-Response-Evaluate type of dialogue around the problems offered on the worksheet. The pattern consisted of Ms. Fleming going over the problems with the students as a whole class. She would move in order through the problems on the worksheet they were currently discussing and for each question, the pattern would be essentially the same." The case study writer explains, "As can be seen from the example, Ms. Fleming asks a question, or reads a question from the sheet to initiate the conversation; next, a student responds to the question with a specific piece of information, either a number, theorem name, or yes/no with little or no emphasis on reasoning or justification; in the next turn of talk Ms. Fleming evaluates the student's response, and then either gives a correct answer if the student answer is incorrect, or poses a new question, which implies the student answer was acceptable." He notes two exceptions to the pattern: (a) a TV game simulation where students are allowed to call on one another for help if they aren't confident of the answer to a question and (b) a small group activity where students worked together on a problem in groups of three or four where, he notes, the groups often took on much the same dynamic as the class overall: students worked on the same problems and usually agreed on an answer, which other students in the group then copied from those who 'got it'.

Which portrait presents the more credible representation of the teacher's practice? What might explain the differences? The difference in the quality of representation and reflection could be attributed, in part, to the differences in format: spontaneous comments in informal conversations and unprepared interviews are unlikely to show the depth of the teacher's considered reflections. And, the written reflections may have been completed with full access to curriculum resources and feedback from colleagues. The handbook in fact encourages collaborative reflection with colleagues. It's also important to keep in mind that the case study occurred at the very end of the year when the teacher was reviewing for the final exam. Does this make the classroom discourse atypical? Given the evidence, it is impossible to know. [We address the issue of ambiguous evidence in more detail in Schutz and Moss (in press).] Whether this would count as a difference that matters depends, in part, on how

portfolio readers cope with the ambiguous evidence in the portfolio.

Additional Comparative Vignettes

We examined all 29 of the comparisons in ELA and Math at the level of detail described and illustrated in the previous two cases. In this section, we present vignettes from five additional comparisons in which the differences we observed did seem to matter in terms of the relevant criteria and raise, we argue, dilemmas that assessment developers and policy makers should consider in the design of assessment systems.

Consistent with the intent of the paper, our vignettes are developed to foreground important differences for a particular comparison; we do not describe, as we have above, the similarities in these comparisons. In the interests of space, we summarize our conclusions with brief illustrations. [We hope the extended examples described above and in the Appendix illustrate the attention to detail that underlies these conclusions.] Each vignette follows a similar pattern: we first characterize the issues and context differences that the vignette raises (so that readers can choose whether to read the vignette) and then we provide brief illustrations of those issues. (Note 14) This section concludes with a brief mention of additional sorts of differences we noted but thought were unlikely to matter in terms of the criteria used. We reserve discussion of the issues the vignettes raise until the final section of the paper where we propose some possible paths for resolution.

Vignette 3: Mr. Richards

In this portfolio/case comparison we see a case in which an English teacher's performance looks substantially different in an honors class than in his third level class. Mr. Richards teaches in what the case study researcher describes as a rural school of about 600 students, 97% of whom are white. Distinguishing among students' placements in the school's tracking system, Mr. Richards indicated, "*Honors kids are chosen because of their work ethic and their intelligence.*" Students in the second level, "*have the work ethic, but they just can't grasp the material. They will eventually, but their work ethic keeps their nose above water.*" For the students in the third level, "*the content is watered down.*" While the portrait of the honors class is relatively consistent across the two methodologies, the case study highlights how it is that his beliefs, as well as institutional tracking, seem to shape his practice with students in different tracks. [The original comparison was prepared by LeeAnn Sutherland.]

Only the honors class is represented in the portfolio as they complete a poetry unit. The case study writer observed the honors class as well as a third level class. Mr. Richards explained that the poems for the third level are not as difficult; they use a narrative, abridged version of the literature selection from their textbook (honors classes read the unabridged version); he uses the textbook much more; he has different expectations of students' writing (the focal "*correction areas*" are different). Mr. Richard's rationale for his choices of texts, for the activities he employs to engage students with literature, and for his implementation of a writing process appear to differ in terms of his understanding of their level of ability.

As the case study writer describes it, both honors and third level students readthe same novel at different times during the school year, but Mr. Richards assessed their interpretive needs and abilities differently. The goal for third level students was more "*the story*" and "*trying to pick out the basic elements [such as] plot, theme*." He believed that honors students, however, "*can go beyond the literal*." Honors students had "*a lot of discussion … a lot of note-taking, explaining the concepts,*" whereas third level students answered primarily lower inference questions on worksheets. Of honors students, Mr. Richards required out-of-class reading and book reports that follow a genre sequence—first fantasy/science fiction, second historical fiction, and the like. Students in the other class did a single book report on a biography or autobiography, and they reported on the book by creating a poster or doing an in-character presentation to the class. They ended the school year with a novel based on a made-for-TV movie which Mr. Richards acknowledged has "*absolutely no literary merit*" but that he chose "*because students like it and because it's reading.*" Students wrote an essay at the end of the unit, a personal narrative that did not require them to make connections with the text itself.

Another example of the difference in opportunities provided to students depending on level was in composition study. Honors students prepared for 10th grade by writing a persuasive essay that included MLA documentation. About writing, Mr. Richards said that honors students would be "*mortified*" to conference with him individually as "*they don't like to be embarrassed*." He typically worked with small groups of these students in the first semester, he said, but did not require "*rewrites*" of them in the second semester because they had already "*mastered the guidelines*" of revision. Mr. Richards writes in the portfolio narrative about two additional, "*authentic*" tasks Honors students would complete— entries for a poetry contest and composing a group poem to be read at graduation.

In contrast to honors students, third level students met with Mr. Richards for individual meetings about their writing. Conferences took place at the front center of the room, facing the class, with the teacher seated at a low table and the student whose paper was being reviewed seated on a high stool next to him. Mr. Richards marked student papers ahead of time so that he would remember what to tell them "*they need to fix.*" He "*counseled kids*" by skimming their papers and calling their attention to each item he had marked as problematic. The case study researcher observed 15 conferences he held over two days. Mr. Richards emphasized form and mechanics in these meetings, including frequent references to spelling, contractions, capitalization, use of second-person pronouns, writing numbers in word form, and the need to include information in its proper place in an essay. Students spoke to answer his questions or to ask for clarification of his suggestions. Following those meetings students were to "*rewrite*," which offered two options. Students could "*rewrite the entire paper and make all the corrections*" or they could rewrite problem words ten times, sentences three times, and in addition, write three things that they learned. Vocabulary study for students in this class consisted of writing definitions, parts of speech, and sentences using each word.

Vignette 4: Mr. Johnson

Here we have a portfolio-portfolio comparison across two different subject areas within mathematics. Both portfolios were generated by a novice middle school mathematics teacher working in a community that he describes as white, suburban, and blue collar. Mr. Johnson works at a large middle school with about 900 students, almost all of whom are native English speakers. The two portfolios present 8th grade math courses; both classes use a popular textbook series. The more advanced of the two is an Algebra I course for "average" students. The unit presented in the portfolio from this class concerns linear relationships, particularly the generation of algebraic equations for lines. The other portfolio is from a Transitions course for "general ability" students. The unit from this course covers statistics, particularly the generation of multiple types of graphs to display data. In a close reading of the two portfolios, important differences emerge relating to the use of 'real world' applications in classroom tasks, to modes of final assessment, and to the role of the teacher in classroom activities. [The original comparison was prepared by Jon Star.]

The first category of difference concerns Mr. Johnson's use of real world examples and concrete materials. Connections to real world examples play a very prominent role in the tasks in the Transitions portfolio. Several of the lessons in this unit begin with students collecting data that is subsequently made into a chart. For example, students count Fruit Loops cereal pieces to determine which colors occur most frequently; they work with box scores of a basketball game; and they cut paper plates in their examination of pie charts. In contrast, context plays little or no role in the Algebra portfolio. Students work exclusively with symbolic equations of lines: these equations are never given any referent or context, nor are any real-life situations embodying linear relationships introduced in class.

A second salient difference concerns the way Mr. Johnson assesses students at the end of the portfolio units. In the Algebra portfolio, the teacher assesses students in a traditional manner -- using a written test, administered in a single class period. Students are asked to complete 23 problems, all clustered around the execution of procedures (finding an equation of a line given a point and a slope, finding the equation of a line given two points, and converting a line from point-slope form to general form). There is significant repetition: for example, clusters of four or five problems look identical, with only the numbers changed from one to the next. In contrast, the final assessment in the Transition portfolio requires students (in groups) to collect data, construct graphs, and give an oral presentation. This assessment takes several days to complete; students are assessed on the quality and accuracy of their graphs and on their oral presentations. At the conclusion of this assessment, students meet individually with the teacher to discuss their grade.

A third difference concerns the role Mr. Johnson appears to take in conducting classroom activities. In the Transitions class, the teacher seems to view his role as one of a background guide; his actions and his commentary consistently indicate that his goal is to largely remove himself from classroom activity. For example, in one lesson plan, he writes that he plans to "*step into the background and let students proceed with their work on their own*." In another lesson, the teacher makes an explicit attempt to re-direct questions posed to

him back to students (and he subsequently reflects that he was very happy with the results). In general, almost all of the Mr. Johnson's lessons in this portfolio consist of students being given a worksheet or an activity to do in groups; the teacher spends much of each class in the background, circulating from group to group and answering students' questions when they arise. In contrast, the teacher portrayed in the Algebra portfolio is much more directly involved in student activity. Almost all lessons in the Algebra portfolio involve the teacher conducting a recitation: standing in front of the classroom, he demonstrates a procedure, asks frequent questions of the class to guide him through his demonstration, and then offers problems that the class should do for practice. Although students are involved in these recitations via the teacher's questions, the teacher is largely controlling the activity and problem-solving that occurs in most classes. The teacher writes that he views the recitation style of instruction as appropriate for the more advanced Algebra class but less so for the low-achieving students of the Transitions class: "*Low achievers and behavior disorder students could not stand more than ten minutes of lecture... The style that works best for them is more of an activity based learning.*"

Vignette 5: Mrs. Martin

This ELA portfolio-case comparison raises a complex chain of issues: (a) we see the same class at two different points in time engaging in substantially different learning activities; (b) we learn from the case that some practices illustrated in the portfolio were not consistent with the teacher's practice, were undertaken because the portfolio handbook prompted them, and were not consistent with what she believed were her students' needs; and (c) this then causes us to question the teacher's judgments about her students' capabilities. [The original comparison was prepared by LeeAnn Sutherland.]

The teacher in this portfolio/case comparison works at a school characterized by the case study writer as a "large inner city school." The students who attend this school are predominantly Hispanic from poor and working class families. For this teacher, Mrs. Martin, both the portfolio and the case study are based on a 9th grade Writing Enrichment course. The course was developed for students in a "transitional program" who are "*too old to be in Middle School*" at 15-16 years of age, but are "*earmarked as an at-risk group.*" There are 13 students in the class, 10 of whom are bilingual; 3 are identified as special education students. The portfolio literature exhibit is comprised of one 4-page short story and one poem which is integrated with the writing exhibit. The case study focuses on a drama unit with a two-page play from an adolescents' literary magazine as the primary text. The case study writer observed two sections of the same writing enrichment course.

The texts used in the literature section of the portfolio were selected to focus on the theme of strong, courageous women. The teacher characterizes her goals as helping students see the connection among these pieces of literature and their own lives, wanting them to "*see the potential within themselves*". The lessons focusing on these texts, across more than 10 of the 15 days represented in the portfolio, take students through a series of activities including the completion of several charts focusing on elements of literature such as character, theme, and imagery, a 2-paragraph "mini-essay" describing one of

the characters, and a culminating essay in which students write three paragraphs which compare a character from the short story with a character from the poem. The comparison writer notes that the time spent on these brief and straightforward texts seems excessive and that students have little opportunity to develop their own ideas. The teacher's reflections suggest that she believes students need this level of support to comprehend the story. She indicates that "*students have difficulty decoding words*" and that each story read in this course begins with an uninterrupted oral reading (by the teacher) that gives the students "*the opportunity to hear the story first, get a basic idea of the plot of the story, and minimize the frustration of difficult vocabulary.*" She notes that "*focusing on a few skills and then building on them, ensures a complete understanding, and more importantly, retention of the lesson. For this group of students, retention is the key.*"

In her portfolio, Mrs. Martin provides commentary and videotape of two students as they conference about essays they have written. Each of the students offers observations to the other, and the author is seen to respond to those observations. They discuss thesis statements, paragraphing, use of examples, and proper citation form including line numbering [for the poem]. They also discuss parts of the essay they found difficult to understand. The comparison writer argues that this is one of the better student-student conferences seen in portfolios, as participants are actively engaged in dialogue about writing. While it is not a substantively rich conference, many of the writing conferences seen on videotape are teacher-directed, or the students speak *to* one another but not *with* one another. The two students seem to "get" the idea of how to engage in a writing conference.

However, when the case study researcher asked the teacher a question from the interview protocol, "Did the portfolio involve things that were not part of your teaching practice?" Mrs. Martin responded: "*I thought the peer editing and peer responses were phony for me because I don't do that yet. The kids were not really ready yet.*" If the teacher coached the girls on how to talk for the camera, then that raises one set of issues. If she did not, but simply asked them to conference, even though she usually does not have them do so, then their relative success in the conference raises questions about the teacher's judgment of students' abilities.

The case study writer saw no writing in response to literature and no process writing during the time of his observations. The only writing he saw involved lists and definitions associated with vocabulary words. Asked whether what the case study writer had observed over the three-day period was "typical of your teaching and your classroom," Mrs. Martin indicated that "*the class is a writing enrichment class and most of our time for the whole year was spent on enrichment.*" She stated that previous writing assignments for the course had followed the [state's] test format and students had also written autobiographical essays. Neither source of evidence provides examples of these types of writing.

About literature, Mrs. Martin said that the portfolio requirements—again—did not jibe with her usual practice: "*I think having 7-8 hours of literature did not really fit the curriculum I have for these kids.*" She told the case study writer that these students are not ready for a novel or for the "*7-8 hours of literature"*

required for the portfolio, that they struggle with decoding, and that students needed to read the screenplay twice in order to get it. The case study writer observed the teaching of the magazine play, and he reports that students' oral reading over two days was relatively fluent, and though little attention was paid to students' understanding of the play, students' verbal comments, a question one student asked, their expressive reading, and other verbal cues indicated that they did, indeed, comprehend this particular text as they were reading. Again, this raises questions for us about the teacher's judgment of her students' capabilities and needs—question that the evidence is insufficient to address.

Vignette 6: Mr. Gere

In this vignette, we encounter a teacher who indicates that he chose to use his portfolio as an opportunity to improve certain areas of his teaching. In the portfolio he presents an Algebra I class, where he reports the students reflected a wide range of abilities and dispositions. The case study writer observed the Algebra 1 class and a Pre-Calculus honors class. We learn from the case that Mr. Gere had originally decided to focus the Pre-Calculus honors class for the portfolio but switched to the Algebra I class because he felt the class needed extra attention. In the portfolio, he indicates that he hoped the portfolio would help him focus in on the difficulties he was having and turn them around. [This vignette is based on the comparison originally developed by Pamela Geist.]

While the two portraits of teaching present quite consistent evidence about this teacher's practice, the interpretive commentary that surrounds them leaves the reader with a substantially different impression about the effectiveness of his teaching. Unlike the situation with Ms. Fleming, where her interpretation highlights the strengths in her practice, Mr. Gere focuses, it seems relentlessly, on his concerns about his teaching and what his students are learning. The case study, then, presents the teacher in a far more positive light.

The case study writer, acknowledging the predominance of procedural work and the ongoing focus on manipulating expressions and equations, nevertheless develops a picture of a teacher who encourages students to look at underlying ideas and explore some of the logic associated with working the procedures. The case study writer concludes, "In all six of the classes observed, the students' oral responses, questions, homework, classwork and quizzes indicated that Mr. Gere's expectations were accessible to most students." She notes some differences between Pre-Calculus and Algebra: For example, the case study writer reports that in the algebra class, the pattern was one of fairly routine mechanics; first distributing with algebraic expressions, then factoring algebraic expressions, and finally solving quadratic equations that were factored and set equal to zero. In the pre-calculus class, although the work appeared to be quite mechanical, there was more problem solving involved because of the number of possibilities when finding equations that fit a set of data points. However, she also notes: "As the material developed over the three days, the students played a bigger role in the dialogue, offering their own strategies for finding the equation from a set of data points. Mr. Gere also used open-ended questions effectively: for example, about a quadratic equation, in standard form, students were asked to give its characteristics, in other words, to tell him what they could about this function. The responses were extensive and showed

depth of knowledge about a quadratic." The case study writer creates an overall image of a fairly successful teacher, one who takes his work seriously, is well-liked and respected by his students, and works hard to create a practice that meets his goals and expectations.

The comparison writer notes the similarities between this representation and what she sees in the artifacts of the portfolio. The portfolio artifacts show a similar continuum of difficulty on daily worksheets, quizzes, and tests. Problems begin with simple equations and progress to more complicated ones. Initial tasks focus on the procedural steps to solve problems and move toward using these steps in context. There usually is one task that requires students to explain an idea or the logic underlying steps. In this sense, both reports show that tasks become progressively more difficult because they require that students know more about the different scenarios represented in algebraic equations and how to manipulate more complex expressions and equations. In large group work, Mr. Gere demonstrates procedures and talks students through his logic of the steps. Mr. Gere asks next-step questions of students and students answer Mr. Gere directly. And he effectively and accurately demonstrates the procedures for students, using appropriate mathematical language and notation to demonstrate how a system is solved, and students practice and memorize the procedures eventually making them their own process. Mr. Gere promotes student-to-student discourse in the context of small group work and pairing students together to complete a task. The video evidence illuminates that for the most part students work productively in pairs and small groups explaining to each other how to proceed with a task and compare procedures and answers with each other.

And yet, Mr. Gere's reflective commentary on his practice paints an entirely different image of his success. For instance, talking about the difficulties he faced in facilitating discussion, he writes "*I regretted not soliciting a variety of problem solving methods for this exercise and again bypassed potentially rich mathematical discussion in the interests of time. The decomposition of the problem's solution into discrete steps was worthwhile and helpful, but again lost something due to the more directed discussion that resulted from my sense of time pressure.*" He notes further: "*My responses to students' questions also reflect my impatience such as the response to student A when I don't even let him finish his question before answering. I give him a perfectly accurate and reasonable answer, but the tone of impatience is more damaging in other areas. Another student question is similar in outcome. I quickly give an accurate concise answer to his question but would have benefited the other students with the same misunderstanding by instead redirecting his question to a few of the weaker students to make sure their understanding was solid.*" He worries "*I have begun to recognize that I have slowly adopted more and more of the students' inclination to 'just let me see how to do the problem so I can stop thinking.'*" In fact, he describes what he perceives to be an ongoing decline in students' efforts to succeed in his Algebra I class: "*I know that the effort level has declined precipitously over the past 1 1/2 months in this class, and I worry that I am enabling the very destructive tendencies that are plaguing this class.*"

Thus, there is a running theme across his reflections, one that details the frustrations and disappointments of not being able to change students' attitude.

The image he creates in the portfolio is of a teacher struggling with changing his teaching and at times, there is a sense of hopelessness. Because there is little offered in the portfolio in the way of a rich analysis for how he intends to turn this pattern around, the portfolio writing produces the image of a teacher who sees himself as mostly ineffective and struggling with supporting richer opportunities in the discourse and at the same time offers few ideas for how to change current patterns. In effect, the case study report paints a much more positive image of the discourse patterns, indicating that procedural goals are getting met through the patterns of discourse and at times, especially in Pre-calculus, the discourse supports a deeper and richer investigation into the mathematical ideas. While the comparison writer, perhaps cued by the image in the case study, was able to read behind Mr. Gere's commentary, this portfolio (which was also used in another study involving multiple readers) elicits quite different reactions depending on the weight the reader gives the teacher's negative commentary about himself. Whether this is a difference that would matter depends on whether or not the portfolio readers are willing and able to read behind the teacher's commentary.

Vignette 7: Mrs. Jacobson

In one sense, this portfolio/portfolio comparison provides another example of a teacher whose practices look different in classes she characterizes as comprised of students with different ability levels. In this case, we observe differences in the teacher's demeanor and attitude toward students in the two classes. We also note that her expectations, her explanations for her choices, and her reflections on students' performance in the second portfolio (unlike the primary portfolio) are sometimes framed in terms of cultural and linguistic differences. [The primary comparison was prepared by Laura Haniford, drawing on documents from Steve Koziol, Leah Kirell, and Suzanne Knight.]

This teacher, Mrs. Jacobson, submitted portfolios for two 7th grade classes that are "theoretically heterogeneous" but that are actually grouped, as she reports, based upon the scheduling of math classes. There are 26 students in the first class and 28 students in the second class; there appear to be 2-3 students of color in the first class while students of color are a majority in the second class. The teacher is white. The base text in both classes is a trade novel set during World War II and is part of her department's prescribed curriculum.

Mrs. Jacobson characterizes her students in the first class as "*bright and fun*" and states that her expectations for students reading this novel are that they "*learn the historical and cultural ramifications of World War II. I intended that students examine the personal struggle of the innocent civilians victimized during the war and the incredible strength and courage of the survivors.*" She also states that this particular selection exposed the students to diverse perspectives "*other than the black/white issue which is pervasive at this school.*" In contrast, Mrs. Jacobson characterizes the students in the second class as "*behavior problems*" and her expectations for them are different. She states that she would not have chosen this book for them and that "*the majority of these students cannot--or will not--read it and understand it. These students are intensely committed to being Black or Hispanic and did not relate to the Holocaust…. They love violence and injustice—most kids their age do…. [But]*

This was far too sanitized for them." Mrs. Jacobson also states that she is more concerned that the students in the second class learn the history of WWII as opposed to understanding any elements of plot or character.

The teacher begins each class with a daily oral language (DOL) experience. In class one, this takes many forms – open ended questions about literary terms followed by a discussion of some examples from the novel and from students own experience; brief comprehension questions on the reading; a vocabulary exercise where extra credit is given for making the teacher laugh; a brief review of grammatical terms. In class two, the DOL is consistently a recitation/review of questions on the assigned reading with answers given "*swiftly*" and written answers handed in at the end of the week. Commenting on an interaction in class one, she writes, "*In the future, I might take a hint from this class and compare movies they may have seen with the books we're reading. I always try to relate what we're doing to their own lives, but they like to talk about movies.*" Of class two, she writes "*With this group, I have to lead them with a strong hand, although I try very hard not to tell them what they 'should' say: I want to hear what they want to say, even if it is immature or downright silly.*"

In both classes, students' written response to literature is related to preparing a five paragraph theme and to addressing the state's criteria for a persuasive essay. Beyond this, her stated goals for the first class include learning to appeal to all five senses, to write "*great opening lines*" and to "*engage*" readers; for the second class, her goal is getting students to write "*something--anything.*" In the first class, the primary assignment asks students to take a position on whether they would take in an escaped prisoner of war who came to their home seeking refuge. In the second class, the same primary assignment is given, together with an alternative prompt related to a reading on the Civil Rights movement, because "*They identified with the black students.*"

In class one, the primary assignment is grouped with two others, a personal time narrative and a group diary designed to personalize the story for students; there are no surrounding assignments in class two and students in the second class are not given the opportunities to work with one another that the students in class one are. Samples of writing from each class suggest that students understood the demands of the assignment and could respond to it appropriately. Commenting on her concerns about a student's writing in the first class, she says: He "*does not create an effective visual in the opening paragraph. Also…[he] does not respond to the opposition in his fourth paragraph. He only states that there is another side.*" Of one student in the second class, she writes: "*[He] did not follow the guidelines, either. This child's family speaks Spanish in the home, and he had made great improvements since September. He has learned to skip the lines to make paragraphs, and he is writing sentences, rather than one long sentence.*"

The guidelines for the state's writing tests are the focus of formal writing instruction in both classes and of video segments on writing. In the first class, Mrs. Jacobson's introduction is brief, mainly an overview, and students have an opportunity to look at some samples and to begin working in a writing workshop format on their own essays – students read aloud some of their drafts, they work together in peer editing, the teacher guides the critique of samples,

drawing from the student samples to deal with topics in language use (e.g., using over-used words), and she confers with individual students. On the video, the teacher circulates around the room, talking with individual students about their work. Overall, the interactions appear positive and supportive.

In class two, the teacher guides students through a series of questions about the state's writing test guidelines, seeking responses about what is to go into each paragraph and elaborating on student responses. She moves to a whole class example – on the topic of what if there were no teachers – which begins to generate student responses, although the teacher appears (on video, as well as in her comments) to be frustrated that the students don't seem to understand how to give reasons for the "*other side,*" which she says is required by the state's guidelines. There is no small group work: students are in whole class activity or working on their own; when they are writing, the teacher circulates and has occasional interactions with students.

Based on our observation of the video, the teacher's management in class one appears to be smooth; students move from one type of activity to another and from one arrangement to another with little disruption; the teacher comments that this group is especially active and noisy, although that doesn't appear to be evident from the tapes. In the second class, management issues dominate more of the dialogue. The teacher writes: "*running a discussion with this group is like walking through hip-deep jello. With every remark comes ambient noise and chatter which drowns it out and everything has to be repeated. In fact, as I watched this segment, I was bored just listening to myself repeat the instructions more than 10 or 20 times. Virtually nothing got accomplished.*" In addition to this, Mrs. Jacobson has several extended disagreements with individual students in the second class that are conducted in front of the entire class.

Mrs. Jacobson's reflection about her teaching and about students' learning is not detailed or extensive. With class one, she notes that some of her assignments were too vague and that she was unprepared for how capable her students were, something she would better prepare for in the future. She thinks she will add some drama in the future, because the group would have done well with this kind of reading and activity. With class two, she notes that she was not particularly effective as a teacher, but attributes this primarily to being required to teach an inappropriate text and having to follow a district mandate that doesn't fit the students. She notes, "Sadly, any of these students' *real* problem is behavior. If they would listen, if they wanted to produce, they could. Peer pressure and stress at home makes it nearly impossible for them to succeed. Patience and in-class time to do their work does increase their chances of doing acceptable work."

Additional Differences

Substantive differences existed in all the comparisons, as we would expect in any dynamic teaching situation. It's important to note, however, that in the majority of cases examined in math and ELA, we found that the second data source elaborated but did not overturn our general impression of the quality of the teacher's performance with respect to the relevant criteria. In some cases,

we simply saw the same practices instantiated in a different content; in some cases we saw somewhat different practices that, taken together, presented a coherent portrait across the two (e.g., Ms. Bertram) or clarified an ambiguity in the original portfolio (e.g., Ms. Fleming); in some cases we saw differences similar to those we represented here but not so substantial as to overturn our judgment of the primary portrait. Portfolios that contained inconsistent evidence--in which the artifacts did not fully support the teacher's descriptions, as with Ms. Fleming--complicate the question of portfolio generalizability with the problem of interpreting the initial portrait. [We discuss issues of portfolio evaluation elsewhere (Schutz and Moss, in press).] In some cases, we learned things about the teacher in the case, which may not have been relevant to the criteria, but which shaped our judgment of the teacher. For instance, in one case (a likely "conditional" score), the case study researcher observed multiple situations of conflict, at least one potentially violent, during and outside of class that the teacher skillfully resolved. In fact, we often learned about the teacher's relationship to, rapport with, and work with students outside of class. We also learned about numerous factors that influenced the teachers' performances that would not likely be mentioned in the portfolio or illuminated by the criteria if they were: the presence or lack of a coherent curriculum and/or text that is consistent with the standards; the presence or lack of a supportive mentor in the teacher's subject area; large differences in professional development opportunities and opportunities for collaborative work with colleagues; differences in resources available to prepare the portfolio (including release time and access to video equipment for multiple days). Of course, whether and how these factors influencing a teacher's performance should or even could be fairly taken into account in this assessment is an open question.

Conclusions

As we indicated in the introduction, our goal is to use these comparisons to illuminate issues for assessment developers to consider in designing assessment systems. Consequently, our analysis was disconfirmatory: It was not intended to document consistency but rather to highlight the kinds of differences that can occur across different representations of a teacher's practice and that point to potential problems with implicit assumptions about generalizability. We want to caution readers against drawing conclusions about the typicality of our comparisons. There is no way to know how these volunteers--teachers who were willing to complete a second portfolio or to allow an observer into their classrooms for 3-5 days--might differ from the larger population of beginning teachers. However, the dilemmas we have found--which would not be illuminated in data that are routinely collected--highlight important issues for educators, assessment developers, psychometricians, and policy makers to consider.

We begin our conclusions with a review of the kinds of differences that seem likely to matter (that is, likely to result in different performance levels) in terms of the relevant criteria. Then we return to the useful concepts of generalizability with which the study was framed. What is the assessment domain (or "universe") to which we can safely generalize? What is the (larger) outcome domain about which we can reasonably draw inferences supported with logical arguments and intermittent empirical studies? How consistent are these

domains with the domain implied in the decision about licensure? We close with some more speculative thoughts about the nature of assessment systems (and theoretical resources) that might support well warranted decisions about teaching performance.

What Have We Learned about the Generalizability of Teaching Portfolios?

The comparisons in this study begin to address an important gap in our understanding of the generalizability of portfolio assessments of teaching and, perhaps, of performance assessments of teaching more generally. Taken together, these vignettes raise a number of concerns, some of which relate directly to the topic of generalizability and some of which spill over into concerns about validity and ethics.

In the small set of comparisons we've examined here, it is very clear that *context matters*. We've shown differences in performance across classes that differ in (perceived and/or institutionally designated) ability level of students, in subject matter taught, and in cultural background of students. For instance, in the case of Mr. Johnson, we saw differences in performance across two subject matter domains: statistics and linear equations in algebra. Perhaps it is easier for novice teachers to develop "rich and challenging" tasks that foster "connections" and "reflect students' interests, styles and experiences" in some domains than in others. We've presented two clear examples of differences in performance across classes that differ in perceived ability level (Mr. Richards and Mrs. Jacobson). And we found other cases (not described here) in which the differences were apparent but far more subtle (as might be seen in the difference between the Pre-Calculus and Algebra classes of Mr. Gere). In Mrs. Jacobson's case, perceived differences in ability were coupled with differences in the cultural background of her students. The differences in performance here are more troubling because of the teacher's apparent attitude toward the students and tendency to seek explanations of their performance outside her practice, in district requirements and in their perceived needs as members of different cultural groups. The rubric has no place for descriptions of teachers' expectations and, indeed, if it did, it would be easy to coach a teacher to eliminate problematic language from her text.

That context matters will come as no surprise to those who study classroom teaching or performance assessment. There are complex and dynamic relationships among teachers' social backgrounds and experiences; their expectations, values, and beliefs; their classroom practices; their students' (inter)actions; and the larger social and institutional structures in which they live and work (Gallego, Cole, and the Laboratory of Human Cognition, 2002; Knapp and Woolverton, 1995; McLaughlin and Little, 1993; McLaughlin, Talbert, Bascia, 1990; McNeill, 1983; Stodolsky and Grossman, 1995). Research in performance assessment more generally, with tasks that are far narrower in scope than those represented in teaching portfolios, shows us that different people perform differently on different tasks (the person x task interaction, in terms of generalizability theory) which necessarily confound the construct of interest with variations in the context in which it is preformed. A recent review of approaches to performance assessment in health professions (Swanson et al., 1995) leads to similar conclusions about the difficulty of generalizing across the

contexts presented by different tasks. “Regardless of the assessment method used, performance in one context (typically, a patient case) does not predict performance in other contexts very well” (Swanson, Norman, and Linn, 1995, p. 8). The social context of a classroom seems even more complex than that of a health professional-patient relationship. While both are certainly equally embedded in societal and institutional structures, the classroom involves dynamic relationships among as many as 30 – 35 individuals, each with their own cultural/personal backgrounds that vary in ways we can’t predict. Gallego and colleagues (2002) argued that “every continuing social group develops a culture and a body of social relations that are peculiar and common to its members…. Hence,… we can expect that every classroom will develop its own variant” (p. 992).

Two recent reviews of assessments of teaching (NRC, 2001b; Porter, Youngs, and Odden, 2003) both raised concerns about the lack of evidence of teaching performance across differing classroom contexts, and our observations support those concerns. It is hard to imagine, however, how a single assessment program could adequately (and fairly) address those concerns. One could ask for samples of teachers' performance in different classroom contexts, as we tried to do, and yet the variations available within teachers’ yearly class loads vary quite substantially from teacher to teacher, and all are considerably narrower than the range of classes and school contexts in which they are licensed to teach. (Note 15) One could imagine other kinds of assessments in which teachers are presented with cases from a range of classroom contexts, and this might provide some relevant evidence; however, asking teachers to plan or evaluate activities for students with whom they have little experience would raise other kinds of validity questions. And the experience in health-related professions with these sorts of simulations suggests that questions of generalizability are likely to remain. There are no straightforward solutions.

The case of Ms. Martin raises a second issue directly relevant to generalizability. Here we find a teacher who perceives that she is in the position of being required to show evidence of a performance that is outside of her routine teaching practice. Does that suggest the portfolio guidelines were too directive or restrictive? Experience with National Board assessments has led developers to conclude that it is important for teachers to understand what is valued in the assessment; being explicit about expectations, within the bounds of construct relevance, is considered important for validity and fairness (Pearlman, in press a, b), and INTASC has emulated their practice. Clearly, portfolio assessments of this sort do not support conclusions about what is typical. What we learn with a “passing” portfolio is whether a teacher and a group of her students can engage in a particular kind of practice and reflection in at least one instance. Teachers may, of course, make choices that are not in their best interests, as was the case with Mr. Gere, who chose a class with which he was struggling and then emphasized his shortcomings. While this is commendable and productive for a professional development activity, it is less than strategic for a high-stakes assessment. Careful instructions to candidates, and examples of successful portfolios, will be important in helping teachers demonstrate the strength of their practice with respect to the standards. It is important to recognize, however, that not all candidates will have

commensurate opportunities to illustrate their practice. Assessors should try to make sure that teachers have the human and material resources they need, including adequate time, access to competent mentors, and access to audiovisual services. Of course, assessors cannot control teachers' work assignments or the schools in which they work. We have to recognize that these factors influence the extent to which teachers can demonstrate a performance consistent with higher scores and design a system that is appropriately skeptical of the validity of its conclusions about individual teachers.

Not surprisingly, differences across methods used here also played a role, with different methods being more or less adequate in providing evidence relevant to different criteria (as we discussed above under structural differences). The portfolio typically offered the teachers in our comparison a better opportunity to explain their choices and reflect thoughtfully on their teaching (although with a skillful interviewer, one could imagine the opposite for some teachers who are uncomfortable with writing); the case study provided more evidence (six full classes vs. brief videotaped segments from two featured lessons) that allowed stronger inferences about the pattern of discourse in class. Of course, either method could be revised to better address these concerns. If we want to draw conclusions about patterns of classroom discourse, having access to two lessons may be insufficient, especially if they do not support the written description in the portfolio. Clearly, more research about this would be most beneficial.

While criteria were not varied in this study (and are typically considered fixed), there are clearly many different ways to instantiate the INTASC principles in specific criteria tied to available evidence. Consider, for instance, the following two principles taken from the ten INTASC (1992) principles on which the subject specific standards are based:

> Principle #2: *The teacher understands how children learn and develop, and can provide learning opportunities that support their intellectual, social and personal development.*
>
> Principle #5: *The teacher uses an understanding of individual and group motivation and behavior to create a learning environment that encourages positive social interaction, active engagement in learning, and self-motivation. (p. 16)*

The portfolio assessment situates evidence and criteria relevant to these principles within particular subject matter contexts where particular approaches to learning are privileged. Alternatively, assessment developers could, as PRAXIS III assessments do, frame criteria and evidence more generally. Consider the following criteria drawn from PRAXIS III Domain B:

> B1: Creating a climate that promotes fairness
>
> B2: Establishing and maintaining rapport with students
>
> B3: Communicating challenging learning expectations to each student

B4: Establishing and maintaining consistent standards of classroom behavior

B5: Making the physical environment as safe and conducive to learning as possible

(Dwyer, 1998, pp. 21-22)

When we have asked INTASC portfolio readers what they would like to attend to that isn't addressed in the rubric, among the issues that repeatedly arise are teachers' relationships with their students and their classroom management. If these criteria are given some or substantial weight in a compensatory assessment, a number of teachers are likely to improve their scores. In fact, we saw one teacher, working in a large urban school, whose ELA performance tended to emphasize form over meaning and single correct interpretations—consistent with the lowest performance level--and yet the case study researcher saw multiple examples of the teacher handling potentially violent conflict among students in ways that successfully diffused the incident. Moreover, the teacher's reflections on this incident were insightful; he commented, for example, on how he learned who he could touch in violent situations. This observation is not a criticism of the INTASC criteria and standards. As Pearlman (in press) points out, every assessment system has to decide what it values and then make those values clear to candidates. Contra Brennan (2001), treating rubrics as fixed seems the reasonable choice although it's important to acknowledge that changes in the rubric will likely result in some changes in who passes and fails (Moss and Schutz, 1999, 2001).

While it is beyond the scope of this paper to address, our comparisons also raise issues relevant to portfolio readers' evaluation of the specific evidence contained in the portfolio. As the case of Ms. Fleming illustrates, portfolios sometimes contain conflicting evidence, especially artifacts that do not fully support the written representation. As we argue elsewhere (Schutz and Moss, in press), portfolios like these can support different interpretations depending on how readers choose to weigh the evidence. Based on recorded dialogue and extended interviews, we show how readers who clearly value the same criteria and describe the same evidence nevertheless construct a different story about the teacher's practice given the evidence in the portfolio. How should an assessment system address ambiguous evidence like this? We return to this issue as we discuss implications (and in Schutz and Moss, in press).

The meaning of central terms like "discussion," "problem solving," "inquiry," "reasoning," and so on is also at issue. If teachers and readers hold different meanings for these terms—attach them to different actions/interactions—then this can affect readers' understanding of classroom practice (for which there may be no accompanying evidence) and/or readers' evaluation of the accuracy of teachers' reflections. Further, the issue of slant in the representation of performance cannot be ignored. Comparisons among Mr. Gere's reflections and those of the case study and comparison writer illustrate the problem, as does a comparison between Mr. Gere's and Ms. Fleming's reflections on their practice. Mr. Gere must depend on the tenacity of portfolio readers in finding the teaching performance behind the negative slant of the reflection. While portfolio readers can likely be trained to score situations like these reliably, especially with

analytic rubrics that might leave the weighting and combining of evidence to the predetermined algorithm, it does not make the ambiguity go away, rather it simply masks the problem behind consistent scores that are unlikely to be challenged with routine procedures.

Many of the vignettes raise an important issue that spills over the bounds of generalizability: how to evaluate the appropriateness of a teacher's practices--the extent to which they are "challenging" for students--based upon the evidence contained in the portfolio. Portfolio readers' understanding of students' interests, needs, and capabilities depends exclusively on the teachers' characterizations and the few artifacts contained in the portfolio (plus readers' own mostly unarticulated experience with similar students). How does a reader decide whether a teacher's practices are appropriate for her students ***or*** reflective of inappropriately low expectations perhaps resulting from ability groupings imposed by the school? One answer to this question, often heard in committee meetings, is that teachers' practices should be evaluated in light of their justification of their choices. However, in our experience, writing evidence-based justifications or reflections is difficult for many beginning teachers. If the quality of the reflection is more crucial to evaluating particular kinds of performances, then some teachers may be differentially disadvantaged. This leads to the disturbing question of whether it is easier for a teacher to receive a higher score with some classes of students than with others. We have seen a number of examples in which portfolio readers' beliefs about the appropriateness of different activities lead to different scores on the same portfolio. Decisions about appropriateness are often underdetermined by the evidence in the portfolio. One could imagine guidelines that ask the teacher to provide more evidence of students' capabilities. In fact, one review panel, noting a similar problem, asked for evidence of students' work from more students and over a much longer period of time. While this appears to be an appealing solution, it risks making the portfolio unmanageable for beginning teachers who almost invariably report on the time-consuming nature of the task. Again, there are no straightforward solutions to the problem.

What Can We Conclude about the Generalizability of Teaching Portfolio Assessments?

How might the issues we've raised be addressed within the bounds of the theoretical resources on generalizability that framed the paper? Returning to the questions with which we began this section: What is the assessment domain (or "universe") to which we can safely generalize? What is the (larger) outcome domain about which we can reasonably draw inferences supported with logical arguments and intermittent empirical studies? How consistent are these domains with the domain implied in the decision about licensure? Our answers are speculative, based on the evidence provided here and the existing literature on performance assessment of teaching.

What is the assessment domain (or "universe") to which we can safely generalize?

Readers can certainly be trained to achieve sufficiently reliable scores for

individual decisions, as the National Board continues to demonstrate. Our experience suggests the importance and feasibility of preparing readers to forthrightly acknowledge problems of ambiguity in the portfolio evidence. If we were rewriting the guiding questions and scoring criteria as a result of this experience, we would structure them to encourage careful triangulation across different sources of evidence: first, so that higher level inferences could be explicitly built up from more descriptive inferences (as we illustrated in our comparison methodology) and second, to ferret out disjunctions in the evidence that might call inferences about performance levels into question. Assuming that this is already enacted informally as part of readers' training, then it could be further supported by the formal procedures and documentation through which they record their evaluations. The portfolios so identified could be sent for additional review and possibly for additional evidence from the beginning teacher. While most large-scale assessment systems are set up to deal with unscorable responses, responses on which readers disagree, or responses that are flagged as atypical in some way (Wilson and Case, 1997; Engelhard, 1994, 2002), we imagine that what we are suggesting might well lead to a larger proportion of responses being identified as needing additional attention, which will, in turn, increase the cost of the system. Having additional evidence of patterns of classroom discourse and of students' learning would be useful and might reduce the number of portfolios that need additional review and/or evidence. Of course, this would increase the time it takes to evaluate each portfolio. It might, however, be possible to develop a multi-stage evaluation system that only examines the additional evidence (beyond the featured lessons, for instance) when questions arise. That said, it is important to note that with the portfolio evidence alone, we have no idea what additional factors enabled or constrained the performance, and which of those we would consider within and outside the construct-relevant bounds of an appropriate resource. If the portfolio asked for such information, it is not clear how it could be corroborated or fairly taken into account.

If we know that a teacher and her students *can* demonstrate certain kinds of performances on at least one occasion, *how far beyond inferences about the particular portfolio might a well warranted assessment domain expand*? There was nothing in our evidence that suggests it would not be possible to include in the assessment domain what a teacher can do in this class (not just on this occasion) and possibly, what a teacher can do in other classes (perceived as) very much like this one. For those comparisons that involved the same class at different points in time or very similar classes, the only differences we found that seemed to matter could be explained by ambiguous evidence (e.g., disjunction between written portfolio and video) or by knowledge that the teachers felt obliged to demonstrate activities that were not part of their routine practice. By very similar we mean classes that cover the same content and in which teachers' expectations about the students' capabilities also appeared to be the same. We are careful here to limit the inference to what they can do (and whether that is replicable), not to what they typically do. However, research on classroom culture, which suggests it is always dynamic and at least partially unique (e.g., Gallego et al., 2002) does raise red flags about even this assumption that should be empirically investigated. Thus, additional research that checks on these assumptions, perhaps with smaller teaching exhibits, or with interview/observation cycles like those of PRAXIS III, would be important. If

feasible, more extended observations would be useful. The advantage of an INTASC-like assessment and our more extended case studies is that we see how a unit unfolds over a series of lessons. Evidence supporting the generalizability of what teacher can do may well need to include such a series of lessons. Thus, we speculate it is possible to build a logical argument, that should be buttressed with periodic empirical studies, of the extent to which we can generalize to an assessment domain that includes classes, subject matter, and students like these. It may be necessary to build multiple assessment opportunities into the assessment system itself to flag candidates whose performance is not consistent. Whether these differences should be treated at the first (reliability) or second (transfer) level of generalization described above is an open empirical question and then a matter of judgment.

Beyond this, our evidence, taken together with the general lack of positive evidence that might overturn it, suggests that we cannot extend the well-warranted assessment domain to different classes within a teacher's regular teaching assignments. Many of the differences we've encountered suggest variations that may only be supportable at the second level of generalizability, as a matter of transfer, if there. Our limited, case based qualitative evidence certainly supports concerns that are raised about score generalizability that takes task sampling (which always involves some differences in context) into account and, indeed, suggests that the problems with portfolio assessments of teaching may be much worse because the variations in context are so complex. We simply do not know how a teacher working with an honors class might perform with a class designated as remedial or how a teacher working with a statistics class might perform if the content were substantially different. Here additional research is very much needed—not only to help appropriately limit inferences about teaching performance but also to understand what changes in a teacher's work context should involve additional professional support. Clearly, the domain implied in the decision to license a teacher, typically constrained only by general subject (or subjects for elementary teachers) and grade or age levels, is greater than can possibly be empirically examined and may, at best, involve weak assumptions and negative arguments (not yet disconfirmed) of the sort that Kane and colleagues (1999) describe.

Given the limited assessment domain our study suggests is likely supportable, is it worth mounting a portfolio assessment of teaching? We do believe that the answer is still yes: If, given adequate time and resources, a teacher is not able to demonstrate in one instance a passing performance, then it makes sense to require additional opportunities for professional development and further demonstration of competence before granting a regular license. This is information about teaching performance that would not be available to a state using only written tests. But is there more we can expect from a performance assessment system?

Returning to First Principles

How does a state education authority, charged with ensuring the competence of the teaching force, undertake that task in a well warranted way? A recent NRC report on *Testing Teacher Candidates* raised concerns about licensure

decisions based only on tests of basic skills and content knowledge (which, themselves, the report notes, have only limited validity evidence) (Note 16) (NRC, 2001b). The authors of the report called, in addition, for:

> Research and development of broad-based indicators of teacher competence, not limited to test-based evidence, should be undertaken; indicators should include assessments of teaching performance in the classroom, of candidates' ability to work effectively with students with diverse learning needs and cultural backgrounds and in a variety of settings, and of competencies that more directly relate to student learning. (p. 172).

Given our conclusions from the previous section and the existing literature on performance assessment, this is a tall order. How can information about these sorts of performances be reasonably taken into account by distant users?

When distant users have access to a classroom portfolio like the one we studied here, they certainly know something more about a teacher's practice that they did before. The question is how to use that information or rather what kind of system can feasibly be developed to support a valid and ethical use of that information. If we theorize this problem within the bounds of conventional approaches to generalizability, then our choices for how to improve the assessment system are limited. As Mislevy and colleagues (2002) noted "compromises in theory and methods … result when we have to gather data to meet the constraints of specific models" (p. 49; see also NRC, 2001a): "When unexamined standard operating procedures fall short, it is often worth the effort to return to first principles" (Mislevy et al., 2003, mss. p. 57).

In addressing this issue, it is important to illuminate the distinction between (a) warranting the validity of the interpretation of a score across individuals with the same score (as psychometrics is positioned to do) and (b) warranting the validity of a consequential decision about an individual (which may be informed by a valid score but typically relies on other/additional kinds of evidence and judgments). That these two sorts of warrant can be different is not a radical suggestion, even within the discourse of educational and psychological measurement. As the testing Standards assert: "In educational settings, a decision or characterization that will have major impact on a student should not be made on the basis of a single test score. Other relevant information should be taken into account if it will enhance the overall validity of the decision" (p. 146). (Note 17) Citing the example of identifying students with special needs, the authors of the *Standards* note: "It is important, that in addition to test scores, other relevant information (e.g., school record, classroom observation, parent report) is taken into account by the professionals making the decision" (p. 147). And yet, psychometrics has little advice to offer about how to combine such evidence into a well warranted interpretation or decision.

We close, then, by pointing in two somewhat different (and yet potentially complementary) directions for returning to first principles to enhance the validity of high-stakes assessment of teaching competence: (a) toward the flexible use of probability based reasoning illustrated in work of Mislevy, Wilson, and their colleagues (Mislevy et al., 2002, 2003; NRC, 2001a; Wilson and Sloane, 2000; Wilson, 1994) and (b) toward enhancing the capability of local education

authorities, in dialogue with the state, to make well warranted and credible recommendations about individual teachers.

Mislevy, Wilson and colleagues argued that fundamental concepts like validity, reliability, and fairness, are broader than any particular set of methods for addressing them: while "familiar formulas and procedures from test theory" work well with "familiar forms of assessment," (p. 1) they risk constraining new forms of assessment that respond to new developments in our understanding of how people learn. For instance, citing Brennan's association of reliability with replication (2001), they noted that "it is less straightforward to know just what repeating the measurement procedure means if the procedure has several steps that could each be done differently … or if some of the steps can't be repeated at all (if a person learns something by working through a task, a second attempt isn't measuring the same level of knowledge)" (p. 17) (issues Brennan acknowledges). They offered a more general characterization of reliability as "the evidentiary value that a given …body of data would provide for a claim—more specifically, the amount of information for revising belief about an inference" (p. 33).

They cited a number of alternative theoretical models within and beyond psychometrics which taken together enhance our capability to model real world situations in reasoning from evidence to inference.Note 18 They noted further that each of these might be considered a special case of a more encompassing approach to probability based reasoning that would allow mixing of existing models and development of new ones.Note 19 Given the sorts of examples they provided, we imagine that models like these would enable distant users to combine portfolio based evidence with other evidence available about the teacher, including routinely collected evidence from existing tests, and briefer embedded assessments that might be collected during a teachers' preservice and induction years. While none of these could be considered interchangeable or random samples from the same assessment domain (as common approaches to generalizability idealize), each of them would help in decreasing uncertainty about a teacher's competence (or accomplishment). Indeed, a consortium of teacher education institutions in California (Performance Assessment for California Teachers [PACT]) is exploring the use of preservice embedded assessments and induction-year teaching exhibits (similar to the INTASC portfolio exhibits) as an alternative to the state's less-contextualized assessment. Mislevy, speaking about assessment of students' opportunity to learn, envisioned that models can be developed which simultaneously take into account important features of the context in which the assessment occurs (personal communication, 12/28/02). The goal in addressing the qualities of reliability and validity is to increase "the fidelity of probability-based models to real-world situations" (p. 49). While it is beyond the scope of this paper and our collective expertise to proceed much further down this road, we point readers in the direction of these scholars' work to highlight the possibility of developing more flexible models with large-scale centralized forms of assessment.

An alternative (and we argue, complementary) direction for returning to first principles involves enhancing the capability of local authorities (districts, teacher education institutions) to make well warranted (and audited or auditable) recommendations to the state about the readiness of individual teachers to

receive a regular license. Portfolio judgments could be combined with other relevant sorts of evidence only routinely available in the local context. This approach suggests different roles for state and local agencies; and a different use for the portfolio at the state level than at the local level. At the state level, the goal would be to audit the practices and judgments about individual candidates that are made at the local level.

To warrant those decisions, we need to move beyond psychometrics or (frequentistic) probability-based reasoning (Note 20) and look to other epistemological/ethical resources (for instance, in anthropology, hermeneutic philosophy, political philosophy and ethics, and the law). The senior author has turned to hermeneutic philosophy--for reasoning from evidence to inference--as a means of warranting knowledge claims and ethical decisions (Note 21) (Moss, 1994, 1996, 1998, in press; Moss and Schutz, 2001, Moss, Schutz, and Collins, 1998). Practices for developing interpretations across disparate sources of evidence and controlling readers' biases can be found in any number of "qualitative" methods texts (e.g., Erickson, 1986).

Of course, when portfolios are constructed and evaluated within a local educational community, a somewhat different set of threats (and benefits) to validity becomes salient. On the positive side, when candidates have multiple opportunities to demonstrate their learning, capabilities, or accomplishments, the stakes for any one assessment decision are reduced. This is particularly true when support for professional development is provided in between assessment episodes. Similarly, when assessors can seek additional information to help in explaining the observed performance, as is true in many local contexts, the burden placed on interpreting the portfolio evidence is reduced. On the negative side, an ongoing relationship between the candidate and the readers can detract from validity by allowing potentially irrelevant knowledge and commitments to be brought to bear on the conclusions about the candidate's performance. It will be important to bring in outside perspectives to the evaluation process, so that potential disabling biases of readers familiar with the candidate and context (whether favorable or unfavorable to the candidate) can be illuminated and self-consciously considered. Readers designated by the state could be invited to participate in the process in a variety of ways--as members of the initial portfolio review team; as auditors of the decision, written documentation, and supporting evidence produced by the team; or as independent reviewers who consider the portfolio based evidence with no knowledge of the outcome. Although consistency in the conclusions of inside and outside reviewers will enhance the validity of the decision, high levels of consistency may be unlikely because of the differing perspectives and knowledge that the different readers bring. Thus, it becomes the state's role to audit and warrant the *process* at the local level, not necessarily the individual decision. Here, the role of the outsider is to illuminate taken-for-granted practices and perspectives and make them available for critical review by members of the interpretive community so that they may be self-consciously affirmed or revised. In that way, the interpretive community continues to evolve in its ability to make sound judgments. Alverno College (NRC, 2001b; Zeichner, 2000) provides one well-documented model of a local system set up to support professional development and warrant high stakes decisions. Descriptions of other examples of local decision making processes can be found in Porter,

Youngs and Odden (2003), NRC (2001b), and Lyons (1998).

We believe both of these approaches--one focused on the warrant for centralized decisions and the other on the warrant for local decisions--might enhance the way in which evidence of teaching performance can be taken into account in licensure decisions. Each has advantages and disadvantages, resolving some validity problems and raising others. Whichever approach is privileged in a given educational context (that is, whichever approach results in the decision that “counts”), the other approach can (and should) provide an important check on or challenge to the validity of those decisions.

In closing, we concur with the National Research Council's (2001b) call for a wider range of assessment practices than is typically gathered at the state level, including evidence of teaching performance. The work of the National Board and of INTASC and Connecticut suggests that portfolios represent one feasible means for obtaining information about teachers’ classroom performance. More research is needed, however, to unearth potential problems with portfolio or other performance-based interpretations and to provoke debate about solutions. Further, it is important to note that the dilemmas we have raised do not simply reflect technical problems. And the solutions, we believe, are not within the bounds of a single assessment program. Rather, teacher education institutions and the schools and districts within which teachers work must work together to support beginning teachers, especially as they move into new contexts, and to ensure that they are ready to provide a productive learning environment for all of their students.

Table 1
Participating ELA Teachers* and Their Classes

(with achievement levels as characterized by teachers)

Portfolio/Case Comparisons

	Primary Portfolio Class(es)	**Non-Portfolio Class(es)**
Ms. Bertram	6th grade language arts(“heterogeneous”)	6th grade language arts(“heterogeneous”)
Mrs. Carson	8th grade English(“advanced” “best math students”)	8th grade English(“best math students”)
Mr. Koehler	8th grade ELA(“technically heterogeneous” but mostly “upper level” due to math track)	8th grade ELA(“heterogeneous”)
Mrs. Martin	9th grade writing enrichment	9th grade writing enrichment
Mr. Richards	9th Grade honors	9th Grade third level (“content is watered down”)

Mr. Roberts	College English 9 (“average”)	College English 9 (“average”)
Mr. Roosevelt	12th grade Humanities II honors *and*10th grade “general”	10th grade honors
Mr. Turner	11th grade English III (“college level”)	9th grade honors

Portfolio/Portfolio Comparisons

	Primary Portfolio Class	Secondary Portfolio Class
Mrs. Harris	10th grade English II (“heterogeneous, vocational”)	10th grade English II (“heterogeneous, vocational”)
Mrs. Jacobson	7th grade Language Arts (“heterogeneous,” “phase II math”)	7th grade Language Arts (“heterogeneous, “lower ability math”)
Mrs. Marks	9th grade World Literature I (“top level”)	10th grade World Literature II (“average track”)
Ms. Patrick	middle school	middle school
Ms. Phillips	7th grade English (“gifted and talented”)	7th grade English (“B-level”)
Ms. Snyder	7th grade English (“many reading 1 – 2 levels below grade level”)	7th grade English (“many reading 1-2 levels below grade level”)

*All names are pseudonyms.

Table 2
Participating Mathematics Teachers* and Their Classes

(with achievement levels as characterized by teachers)

Portfolio/Case Comparisons

	Primary Portfolio Class(es)	Non-Portfolio Class(es)
Ms. Anderson	Algebra I primarily 9th graders	Accelerated Algebra I all 9th graders
Ms. Fleming	Integrated Geometry 11th graders	Math I remedial math (“failing students”)

Mr. Gere	Algebra I 9th graders	Pre-Calculus predominantly juniors, a few 10th graders.
Mrs. Green	Geometry (academic) 10th, 11th and 12th graders	Integrated Math remedial math (“socially promoted students”)
Mrs. Jones	Geometry (college bound) 10th and 11th graders	Algebra I (college bound) 9th graders
Ms. Rinaldi	General 8th grade mathematics	Accelerated Geometry All 8th grade students (most accelerated math students in the school)
Mr. Skinner	Pre-Calculus (optional 4th year of mathematics) 12th graders	Algebra II 11th and 12th grade (students of varying abilities)
Ms. Weaver	Algebra II (primarily college bound students, some repeating the course) 10th, 11th, and 12th graders	Geometry (majority of the students are college bound, but not honors students) - 10th, 11th, and 12th graders

Portfolio/Portfolio Comparisons

	First Portfolio Class	Second Portfolio Class
Ms. Barnes	Algebra and Geometry 8th graders (general ability)	Algebra 8th graders (most advanced math class offered)
Ms. Eastman	Trigonometry/ Analytic Geometry 11th and 12th graders (general ability, high achieving students)	Consumer Math10th, 11th, and 12th graders (many repeating the course or have previously failed a math course)
Mr. Freeman	Integrated Level I (Algebra) 9th graders (advanced)	Geometry 9th and 10th graders (advanced)
Mr. Johnson	Transition Math 8th graders (general ability)	Algebra 8th graders (average ability)
Ms. Layton	Transition to College Mathematics 12th graders (average ability)	Algebra I – part 2 9th, 10th, 11th, and 12th graders (average ability)
Ms. Schafer	Regular/Inclusion Math 7th grade	Regular Education Mathematics 7th grade
Mr. Sexton	8th grade mathematics(students of varying abilities, most advanced students leave this classroom	7th grade mathematics(students of varying abilities)

	to take algebra)	

*All names are pseudonyms.

Notes

- 1. This study was supported, in part, by a grant from the Spencer Foundation. We gratefully acknowledge their support. We also wish to thank Kevin Basmadjian, Leslie Burns, Leah Kirell, Suzanne Knight, Emily Smith, Michigan State University, and Vilma Mesa, University of Michigan, for their thoughtful contributions to the comparative analyses. The senior author is a member of INTASC'S technical advisory committee (TAC). We gratefully acknowledge comments on an earlier draft from Aaron Schutz, Mark Wilson and from INTASC staff and TAC members: Mary Diez, Jim Gee, Ann Gere, Bob Linn, Jean Miller, David Paradise, and Diana Pullin. Opinions, findings, conclusions, and recommendations expressed are those of the authors and do not necessarily reflect the views of INTASC, its technical advisory committee, or its participating states.
- 2. Kane and colleagues (1999) actually refer to three levels of inference: "inferences from performances to observed scores", "inferences [or 'generalization'] from observed scores to universe scores…which includes performances on tasks similar to (i.e., exchangeable with) those in the assessment"," and "inferences [or 'extrapolation'] from universe scores to target scores" reflecting "a larger, and generally less well-defined domain" where the regulative ideal of random sampling is untenable. Messick's (1994) distinction between task- and construct- based assessments seems to parallel Kane's first two levels of inference. Haertel (1985), too, characterizes three levels of generalizability. He differentiates the outcome domain into that part that can be empirically investigated and that part that involves only weak assumptions.
- 3. Brennan notes a discontinuity between IRT and other measurement models: "It is certainly true that statistics that have a reliability like form can be computed based on an IRT analysis, but it is equally true that almost all such analyses treat items as fixed. This raises important questions about what such statistics mean from the point of view of replications of measurement procedures" (2001, p. 304).
- 4. Brennan also cites scoring rubrics and rater training procedures as potentially relevant sources of variation, although most assessments treat these (within the universe of generalization) as fixed.
- 5. There is a long history in the writing assessment literature of examining relationships between so called direct and indirect methods of assessment, both to document the validity of the multiple choice method and to show that actual samples of writing represent a different construct than what can be examined with multiple choice tests.
- 6. Readers should note that the structure of the National Board assessment is undergoing revision; the description presented here was operative in each of the research studies we describe. See www.nbpts.org for updated information.

- 7. For the six certificates that were operational when their Technical Analysis Report was released (1998), the overall estimate of exercise reliability (across the 10 tasks) ranged from 0.72 – 0.87, with a median of 0.825. This included 4 in-class portfolio exercises, two documented accomplishments portfolio exercises, and six assessment center exercises. For the four in-class portfolio exercises, the reliability ranged from .049 – 0.76, with a median of 0.695. [The one math certificate, adolescence and Young Adulthood/Math received the highest exercise reliabilities and Early Adolescence Generalist, the lowest.] (p. 109). Decision consistency for exercise sampling ranged from 5% - 7% for false negatives (estimated percent of candidates who incorrectly failed) and 6% - 9% for false positives (estimated percent of candidates who incorrectly passed). They note that decisions are more consistent for candidates with scores further from the cut score. Assessor reliability was generally high (.90 - .98) overall and (.85 - .92) on in class portfolio exercises. “Based on these analyses of the technical measurement quality of the six certificates administered in the 1996-97, the assessments fully meet the requirements of the Standards for Educational and Psychological Testing (AERA, APA, NCME, 1985) for validity, reliability, and freedom from bias” (p. 125).
- 8. The use of guiding questions that integrate standards into dimensions directed at a particular teaching performance to produce an interpretive summary was developed by Genette Delandshere, Steve Koziol, Penny Pence, Ray Pecheone, Tony Petrosky, and Bill Thompson in their leadership of one of the first two National Board Assessment development labs. This has informed both the work at INTASC and NBPTS. See, for instance Delandshere and Petrosky (1994); Koziol, Burns and Brass (2003).
- 9. With these independent readings, the rank ordering of the two records of teaching were the same. Given the time consuming nature of the task, we decided it was appropriate to move to a single comparison document and audit described next rather than to have two separate documents produced.
- 10. While the four separate classroom based exercises in the National Board portfolio may (or may not) encompass some of these variations depending on the class(es) the teacher chooses for each exercise, they are thoroughly confounded with task differences and not to the best of our knowledge routinely examined.
- 11. All names are pseudonyms.
- 12. Connecticut's ELA guiding questions and rubrics have been revised since we used them. The current versions are available on the Connecticut Department of Education’s web page at http://www.state.ct.us/sde/dtl/t-a/best/portfolio/rubrics.htm .
- 13. Connecticut's Math guiding questions and rubrics have been revised since we used them. The current versions are available on the Connecticut Department of Education’s web page at http://www.state.ct.us/sde/dtl/t-a/best/portfolio/rubrics.htm.
- 14. Direct quotations from the teachers are in italics.
- 15. The National Board assessments encourage but do not require teachers to choose different classes for different tasks (see guidelines at www.nbtps.org). We have not located any documentation that provides evidence about the ways in which teachers respond to this direction,

whether/how they are considered during scoring, or how these differences might shape the evaluations of their performances.

- 16. "Little information abut the technical soundness of teacher licensure tests appears in the published literature. Little research exists on the extent to which licensure tests identify candidates with the knowledge and skills necessary to be minimally competent beginning teachers" (NRC, 2001b, p. 14).
- 17. Of course, the import of this statement depends on what your conception of validity is.
- 18. These include item response theory models, latent class models, structural equation models, and hierarchical models.
- 19. “In the same conceptual framework and with the same estimation approach, we can carry out probability-based reasoning with all of the models we have discussed. Moreover, we can mix and match components of these models, and create new ones, to produce models that correspond to assessment designs motivated by theory and purpose” (Mislevy et al., 2001, p. 49).
- 20. Kadane and Schum (1996), a resource on which Mislevy (1994) draws, cite subjective versions of probabilistic reasoning that could be used with singular judgments. They use this approach to model the likelihood of potential verdicts in a murder trial. They offer this approach as a supplement, a way of illuminating assumptions behind, but not a replacement to the kinds of human judgments involved in such complex social situations.
- 21. Like psychometrics, hermeneutics characterizes a general approach to the interpretation of human products, expressions, or actions. Important differences between these disciplines lie, in part, in the ways in which the information is combined. Psychometric practices support aggregative strategies for combining information: scores for distinct (ideally independent) pieces of information are (weighted and) aggregated to form an interpretable overall score or grade. Hermeneutics supports 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 and Schrift, 1990; Schmidt, 1995).
- 22. Quotations marks (“ ”) indicate quotations from the beginning teacher. Side-ways carats (<>) indicate quotations from the case study writer. All names are pseudonyms.

References

AERA, APA, & NCME (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Ball, D. L., Gere, A. R. & Moss, P. A. (1998). *Fieldwork Guide for the INTASC Beginning Teacher Case Study Project*. Unpublished manuscript, University of Michigan.

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

Bond, L., Smith, T., Baker, W.K., & Hattie, J.A. (2000). *The certification system of the National Board for Professional Teaching Standards: A construct and consequential validity study*. Greensboro, NC: Center for Educational Research and Evaluation, The University of North Carolina at Greensboro.

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. A. (1987). *Assessing writing skill* (Research Monograph No. 11). New York: College Entrance Examination Board.

Brennan, R. L. (1983). *The elements of generalizability theory*. Iowa City, IA: American College Testing Program.

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, *38*(4), 295-317.

Brennan, R., & Johnson, E. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice*, *14*(4), 9-12, 27.

Crehan, K. D. (2001). An investigation of the validity of scores on locally developed performance measures in a school assessment program. *Educational and Psychological Measurement, 61*(5), 841-848.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer (Ed.), *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, theory and public policy* (pp. 147-171). Urbana: University of Illinois Press.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness." *Educational and Psychological Measurement*, *3*(57), 373-399.

Delandshere, G., & Petrosky, A. R. (1994). Capturing teachers' knowledge. *Educational Researcher*, *23* (5), pp. 11-18.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, *4*(4), 289-303.

Dwyer, C. A. (1998). Psychometrics of Praxis III: Classroom performance assessments. *Journal of Personnel Evaluation in Education*, *12*(2), 163-187.

Educational Testing Service (ETS) (1998). *NBPTS technical analysis report,1996-97 administration*. Southfield, MI: NBPTS

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93-112.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large scale assessment programs for all students* (pp. 261-288). Mahwah, NJ: Erlbaum.

Erickson, F. (1986). Qualitative methods in research on teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (pp. 119-161). New York: Macmillan.

Gallego, M. A., Cole, M., & The Laboratory of Human Cognition (2002). Classroom cultures and cultures in Classrooms. In V. Richardson (Ed.), *Handbook of Research on Teaching* (4th Ed.) (pp. 951-997). Washington: AERA.

Gao, X., & Colton, D. A. (1997). Evaluating measurement precisions of performance assessment with multiple forms, raters, and tasks. In D. A. Colton (Ed.), *Reliability issues with performance assessments: A collection of papers*. Iowa City, American College Testing Program. (ACT Research Report Series 97-3).

Gao, X., Shavelson, R. J., & Baxter, G. P. (1994). Generalizability of large-scale performance

assessments in science: Promises and problems. *Applied Measurement in Education*, *7*(4), 323-342.

Haertel, E. (1985). Construct validity and criterion-referenced testing. *Review of Educational Research*, *55*, 23-46.

Haertel, E. H. & Lorie, W. (in press). Validating Standards-Based Test Score interpretations. *Measurement: Interdisciplinary Research and Perspectives*.

Harris, D. J. (1997). Using reliabilities to make decision. In D. A. Colton (Ed.), *Reliability issues with performance assessments: A collection of papers*. Iowa City, American College Testing Program. (ACT Research Report Series 97-3).

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

Jaeger, R. M. (1998). Evaluating the psychometric qualities of the National Board for Professional Teaching Standards' Assessments: A methodological accounting. *Journal of Personnel Evaluation in Education*, *2*(2), 189-210.

Kadane, J.B., & Schum, D.A. (1996). *A probabilistic analysis of the Sacco and Vanzetti evidence*. New York: Wiley.

Kane, M., Crooks, Terence, & Cohen, Allan (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18*(2), 5-17.

Knapp, M. & S. Woolverton. (1995). Social class and schooling. In James Banks & Cherry A. McGee Banks (Eds.), *Handbook of Research on Multicultural Education* (pp. 548-569). New York: Simon and Schuster.

Klein, S. P., McCaffrey, D., Stecher, B., & Koretz, D. (1995). The reliability of mathematics portfolio scores: Lessons from the Vermont experience. *Applied Measurement in Education*, *8* (3), 243-260.

Koretz, D., McCaffrey, D., Klein, S., Bell, R., & Stecher, B. (1992). *The reliability of scores from the 1992 Vermont Portfolio Assessment Program: Interim report*. Santa Monica, CA: Rand Institute on Education and Training, National Center for Research on Evaluation, Standards, and Student Testing.

Koretz, D., Klein, S., McCaffrey, D., & Stecher, B. (1993). *Interim report: The reliability of Vermont portfolio scores in the 1992-93 school year* (RAND, RP-260). Santa Monica, CA: RAND. (Reprinted from CSE Technical Report 370, Los Angeles, University of California, Center for Research on Evaluation, Standards, and Student Testing, December.)

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (in press). The Vermont Portfolio Assessment Program: Findings and implications. *Educational Measurement: Issues and Practice*, *13*(3), 5-16.

Koretz, D., Stecher, B., Klein, S., McCaffrey, D., & Deibert, E. (1993). *Can portfolios assess student performance and influence instruction? The 1991-92 Vermont experience* (RAND, RP-259). Santa Monica, CA: RAND. (Reprinted from CSE Technical Report 371, Los Angeles, University of California, Center for Research on Evaluation, Standards, and Student Testing, December.)

Koziol, S. M. Jr., Burns, L., & Brass, J (2003). *Four lenses for the analysis of teaching. Supporting beginning teachers' practice*. Working paper, Michigan State University.

Lane, S.. Liu., M., Ankenmann, R. D., & Stone, C. A. (1996). Generalizability and validity of a mathematics performance assessment. *Journal of Educational Measurement*, *33*(1), 71-92.

Linn, R. L. & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice*, *13*(1), 5-8, 15.

Lyons, N. (1998) *With portfolios in hand: validation the new teacher professionalism*. New York: Teachers College Press.

McBee, M. M., & Barnes, L. L. (1998). The generalizability of a performance assessment measuring

achievement in eight-grade mathematics. *Applied Measurement in Education*, *11*(2),179-194.

McLaughlin, M., Talbert, J., & Bascia, N. (Eds.). (1990). *The contexts of teaching in secondary schools: teachers' realities*. New York: Teachers College Press.

McLaughlin, M. & Little, J. W. (Eds.). (1993). *Teachers' work: Individuals, colleagues, and contexts*. New York: Teachers College Press.

McNeil, L. (1983). *Contradictions of control: School structure and school knowledge*. New York: Routledge and K. Paul.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Education Researcher*, *32*(2), 13-23.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241-256.

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

Mislevy, R. J., Almond, R., & Steinberg, L. (2003). On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspectives*, *1* (1), pp. 3-62.

Mislevy, R. J., Wilson, M. R., Ercikan, K., Chdowsky, N. (2002). *Psychometric principles in student assessment*. Los Angeles: CRESST.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, *62* (3), 229-258.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*, *23*(2), 5-12.

Moss, P. A. (1996). Enlarging the dialogue in educational measurement: Voices from interpretive research traditions. *Educational Researcher*, *25* (1),20-28, 43.

Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice*, *17* (2), 5-12.

Moss, P. A. (in press). The meaning and consequences of reliability. *Journal of Educational and Behavioral Statistics.*

Moss, P. A., Rex, L., & Geist, P. (2000a). *Case Study Writing Guide for the INTASC Beginning Teacher Case Study Project*. Unpublished manuscript, University of Michigan.

Moss, P. A., Rex, L., & Geist, P. (2000b). *Fieldwork Guide for the INTASC Beginning Teacher Case Study Project.* Unpublished manuscript, University of Michigan.

Moss, P. A. & Schutz, A. (1999). Risking frankness in educational assessment. *Phi Delta Kappan*, *80*(9), 680-687.

Moss, P. A. & Schutz, A. (2001). Educational standards, assessment, and the search for "consensus". *American Educational Research Journal*, *38* (1), 37-70.

Moss, P. A., Schutz, A. M., & Collins, K. M (1998). An Integrative approach to portfolio evaluation for teacher licensure. *Journal of Personnel Evaluation in Education*, *12*(2),139-161.

Moss, P. A., Schutz, A. M., Haniford, L., Miller, R., & Coggshall, J. (in preparation). *High stakes assessment as ethical decision making*. Unpublished manuscript, University of Michigan.

Myford, C. M. (1993). *Formative studies of Praxis III: Classroom Performance Assessments--An overview. The Praxis Series: Professional Assessments for Beginning Teachers*. Princeton, NJ: Educational Testing Service.

Myford, C. M., & Engelhard, G. (2001). Examining the psychometric quality of the national board for professional teaching standards early childhood/generalist assessment System. *Journal of Personnel Evaluation in Education*, *15*(4), 253-285.

Myford, C. M., & Mislevy, R. J. (1995). *Monitoring and improving a portfolio assessment system* (Center for Performance Assessment Research Report). Princeton, NJ: Educational Testing Service.

National Research Council (2001a). *Knowing what students know: The science and design of educational assessment*. Washington, D.C.: National Academy Press.

National Research Council (2001b). *Testing teacher candidates: The role of licensure tests in improving teacher quality*. Washington, D.C.: National Academy Press.

Nystrand, M., Cohen, A. S., & Martinez, D. (1993). Addressing reliability problems in the portfolio assessment of college writing. *Educational Assessment*, *1*(1), 53-70.

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

Pearlman, M. (in press a). The design architecture of NBPTS certification assessments. In L. Ingvarson (Ed.), *Assessing teachers for professional certification*. Stamford, CT: Jai.

Pearlman, M. (in press b). The evolution of the scoring system for NBPTS assessments. In L. Ingvarson (Ed.), *Assessing teachers for professional certification*. Stamford, CT: Jai.

Porter, A. C., Youngs, P. & Odden, A (2003). Advances in teacher assessments and their uses. In V. Richardson (Ed.), *Handbook on Research on Teaching* (pp. 259-297). Washington, DC: AERA.

Reckase, M. D. (1995). Portfolio assessment: A theoretical estimate of score reliability. *Educational Measurement: Issues and Practice*, *14*(1), 12-14, 31.

Schmidt, L. K, (1995). Introduction: Between certainty and relativism. In. L. K. Schmidt (Ed.), *The specter of relativism: Truth, dialogue, and phronesis in philosophical hermeneutics* (pp. 1-22). Evanston, IL: Northwestern University Press.

Schutz, A & Moss, P. A. (in press), Reasonable decisions in portfolio assessment. *Educational Policy Analysis Archives.*

Shavelson, R. J., Baxter, G. P., Pine, J., Yure, J., Goldman, S.R., Smith, B. (1991). Alternative assessment technologies for large scale science assessment: Instrument of education reform. *School effectiveness and school improvement*, *2*(2), 97-114.

Shavelson, R. J., & Webb, N. W. (1991). *Generalizability theory: A primer*. Newbury Park, CA: SAGE Publications.

Stodolsky, S. & Grossman, P. (1995). The impact of subject matter on curricular activity: An analysis of five academic subjects. *American Educational Research Journal*, *32*(2), 227-49.

Swanson, D., Norman, G. R. & Linn, R. L. (1995). Performance-based assessment: Lessons from the health professions. *Educational Researcher*, *24*(5), 5-11,35.

Wilson, M. (1994). Community of judgment: A teacher-centered approach to educational accountability. In Office of Technology Assessment (Ed.), *Issues in Educational Accountability*. Washington, D.C.: Office of Technology Assessment, United States Congress.

Wilson, M., & Case, H. (1997). *An examination of variation in rater severity over time: A study in rater drift*. Berkeley, CA: Berkeley Evaluation and Assessment Research (BEAR) Center.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education*, *13*(2), 181-208.

Zeichner, K. (2000). Alverno College. In L. Darling Hammond (Ed.), *Studies in excellence in teacher education: Preparation in the undergraduate years*. Washington, DC: American Association of Colleges of Teacher Education

About the Authors

Pamela A. Moss is an Associate Professor in the School of Education at the University of Michigan. Her areas of specialization are at the intersections of educational assessment, validity theory, and interpretive social science. She can be reached at 4220 School of Education, University of Michigan, Ann Arbor, MI 48109-1259 (pamoss@umich.edu).

LeeAnn M. Sutherland is an Assistant Research Scientist at the University of Michigan (lsutherl@umich.edu). Her work focuses on adolescent literacy and identity, particularly as students make sense of school discourse vis-à-vis their everyday experiences.

Laura Haniford is a doctoral candidate in Educational Foundations and Policy at the University of Michigan (lhanifor@umich.edu). She specializes in teacher education, especially multicultural education. Her research focuses on the ways in which classroom discourse influences learning opportunities.

Renee Miller is a doctoral candidate in the School of Education at the University of Michigan (reneelm@umich.edu). She specializes in Science Education and Museum Studies.

David Johnson is a Ph.D. student in Educational Studies at the University of Michigan (djjohnso@umich.edu). His research interests include the influence of government policy on teachers and students and how students make meaning of their state-mandated testing experiences.

Pamela K. Geist is an educational consultant in Denver, CO (pamgeist@TEG-Global.com). She specializes in mathematics education.

Stephen M. Koziol, Jr. is Professor and Chair of the Department of Curriculum and Instruction at the University of Maryland (skoziol@umd.edu). He specializes in English Education, program design and policy in teacher education, and teacher assessment.

Jon R. Star is an Assistant Professor in the College of Education at Michigan State University (jonstar@msu.edu). His research focuses on students[1] learning of middle and secondary school mathematics, particularly the development of mathematical understanding in algebra.

Raymond L. Pecheone is an Academic Research and Program Officer in the School of Education at Stanford University (raymond.pecheone@stanford.edu). He specializes in the design and implementation of complex performance assessment systems. Previously Dr. Pecheone was Bureau Chief of Curriculum, Research, Testing and Assessment for the Connecticut State Department of Education. He was co-director of one of the first assessment development labs for the National Board for Professional Teaching Standards.

Appendix

Excerpts From 70 Page Document Comparing Portfolio And Case For "Ms. Bertram" (prepared by LeeAnn M. Sutherland).

Note: We have emphasized excerpts from the exhibits that focus on writing. Consequently, readers may not find examples of evidence for reading in the conclusions which address writing and reading together. (Note 22)

Portfolio	Case
Teacher's goals for the unit	
[based on T's commentary] The T states that the unit we see in the pf, based on the study of a popular age appropriate novella, has three goals. She writes that Ss will be expected to: • Identify the character traits of the main characters in the novella, • Compose a written response citing which character they felt they were most similar to/could relate to best/liked the best and why, and • Use that reflection as the foundation for creating a simulated journal to records thoughts and reflections upon the events of the plot as viewed through that character's eyes, to be evaluated at the end of the novel by the teacher and a pre-discussed rubric (TR, 2).	*[based on interview notes and fieldnotes]* According to the CSW, <one of this teacher's goals centered around the Ss assessment of their own work …. She asked that Ss review all of the entries written in their notebooks throughout out the entire school year and using selection criteria, determine the piece which would serve as their best entry> (CSW, 25). The 3-day series of lessons observed by the CSW were focused on having Ss select 3 best pieces of their own writing from their notebooks, complete evaluation sheets about each one, exchange with a partner who would name his or her choice for the writer's best piece, revision of that piece of writing, and publication on a web page. This is their choice to say "I have grown as a writer and I have matured. This is the piece out of all of the others that I am proud to call my own." (CSW, FN 12).
Teacher's characterization of students	
Based on T's commentary The T says: "In a previous unit on biography, I had realized that this particular group of Ss was both willing and capable of 'allying' themselves with characters whose gender, life experiences, or age were different than their own, if the Ss perceived a connection based on personality or motivation" (5). Thus she "hoped that Ss would choose to further	*Based on interview notes* The T describes her 1st period class as "the sleeper class" (CSW, 8) and <my little puppies coming in every shape and size and personality. I wish I knew them as people … They don't fight over grades, uncertainties or anything, but rather sit there and intimate, 'Let's just get on with it' > (CSW, 9). She describes her 5/6 period as the "hoopla class," one she calls a <challenge class for me> (CSW, 8).

this exploration—that S diaries might cross over the lines of age and gender perspectives, encouraging discussion and deeper reflection into the theme of the book" (5). She says this class is "fearless in its class discussions, and the whole-class discussion forum works well—rather than a teacher-led review, these discussions often erupt into lives of their own. Students frequently—though politely—challenge one another's ideas" (14, L). The T generalizes about this class, "[Although] this class is, overall, a group very much at ease with higher-order thinking skills, there are a number of more concrete Ss who require questions of a more literal nature" (26).	She describes period 2/3, the portfolio class, as <the group I can always count on. They have the ability to do whatever well> (CSW, 9). They are "a class I never have to worry about. They are wonderful, enthusiastic, we will do anything for you today, Ms. Bertram …Their personalities are like a prism…period 1 tend to be a little variegated and textual …whereas period 2 kids come up and share their lives with me" (CSW, 10). "Period two is the class that I'll toss an idea out to and all of a sudden they're coming up with ideas. Boom, Boom, Boom . . . It's their ideas" (CSW, 11). During the first pre-observation interview, the T talked with the CSW about how she anticipated Ss <[responding] to her instructional delivery the following day.> She differentiated between 1st period and 2/3rd period (the pf class): "Period one students will find evaluating themselves as difficult. They think they know what to look for, but … these children often require lots more modeling before they fully understand how to evaluate. They will also struggle with the technical aspects of this assignment. Some are still struggling with using laptops finding it difficult to SAVE their work or to save their work to a disk. Picking the entry itself will be very easy for these children, but the chosen entry may not even be their best piece. However, period two students will find the lesson a challenge, yet a breeze. Most of them are technologically literate and experience few technical difficulties using the laptops. They will be quite enthusiastic about the assignment as well as honest. These children know what to look for when selecting their best piece of work and are most capable of knowing the right answers whatever the lesson."
Chronological summary of activities	

[based on T's daily log] Session 1 SSR—Sustained Silent Reading (10 min). The T introduced the fantasy unit by having Ss brainstorm as a class “What elements would we expect to find in a fantasy novel?” The class then discussed conflict in fantasy and the problem-solving role of the hero. In small groups, Ss brainstormed words and phrases that describe a hero, and they used crayons to sketch an image of a hero. These responses were shared with the whole group. The T encouraged discussion by asking Ss to elaborate; for example, when a S offered: “The hero may not have planned to be a hero,” the T asked the class, “What does this mean?” to encourage discussion. S homework was to write one paragraph describing “The Perfect Hero” (L, 6-7). Session 2 SSR (10 min.) The class began with Ss sharing their homework about the perfect hero aloud. Afterward, the T guided Ss into identifying those characteristics they have listed as “traits,” and talking about differences between physical and nonphysical character traits. Ss then played a game (10 min.) in which they circulated around the room to get their peers’ signatures on a handout that asked for information about both the physical and the nonphysical traits of their classmates. The worksheet required Ss to learn about 24 topics including who owns a kitten, who sings in the shower, and who has freckles?	[based on fieldnotes] Day One Class begins with Ss writing for about 10 minutes in their notebooks. <She then begins the lesson with, “Who can start the review of yesterday’s lesson?” A student responds that they did a notebook share. She continues with, “And what’s the purpose of sharing your notebook entries with others?”> (CSW, FN 15). T does a power point presentation to walk Ss through the steps of choosing their “best piece” for polishing. Discussion ensues with Ss volunteering strategies for choosing a best piece, reasons why they might consider one “best” and why they might “pass” on others. Power point presentation includes a sample entry which is discussed as a whole class. Homework is a “Nomination Ballot” which Ss are to use to review the 3 entries they had chosen for the previous night’s homework. <”You’re going to check off the strong points, not so strong points and comment on whey these three entries could be the entry of the year”> (CSW, FN 20). Because the T realized after 1st period that the evaluation handout was going to be homework, she took more time in the 2nd period (pf class) such as <elaborating more on the concepts> as the class worked through the power point presentation and discussion (CSW, 28). Day Two Class began with 10-minute writing in notebooks followed by a discussion of what Ss found easy and found difficult in last night’s homework. The T noticed that several Ss had incomplete assignments, so she reviewed the concept of “reflection” on the best piece, answering “Why” it could be a winner. Day Three

(ART, 9A). When finished, Ss shared what they had learned about their classmates and discussed “which character traits convey the most information about a person or character” (L, 9). […sessions 3-8 described….] Session 9 T read-aloud continued (10 min.) Following this, Ss received instructions and did individual seatwork in preparation for the following day’s Goldfish Bowl activity. Ss were to answer 4 questions on a handout, and the T circulated to answer questions and keep students “on task.” Questions include: “Do the main characters fit the criteria of a hero—at least, the hero we discussed in our first lesson? Why or why not?” Ss were also asked a prediction question, a question about changing the novel’s point of view, and a question relating faults as sometimes helpful to a particular character in the novel—to how faults could be helpful in real life (23a). The last 15 minutes of the period were devoted to small group discussion, assigning a single to each member, revising individual responses based on other group members’ responses, and preparing to share those the following day (22-23, 23a, L). Session 10—the videotaped session (No oral reading.) Spent the entire class period on the Goldfish Bowl activity. A “Reflection Sheet” (ART, 25A) asked Ss to respond to 3 questions about each of the larger questions discussed in the	Class began with 10-min writing in notebooks and review of what they have done thus far in regard to the notebooks. The next step is revising, which the T walked Ss through using a Power Point presentation. They were also given a handout to guide them. A difference was noted in the way homework was assigned to both classes. Period one was to finish the teacher created ditto, while period two was to complete not only the ditto but revisions on their best entry of the year as well. The teacher reported that it was a mistake on her part, realizing that period one should have finished their revisions as well

bowl. The worksheet asked, “Which part of the discussion most closely matched your own answer? How?” and “Did you disagree with or not understand some part of the discussion? Which part? Why?” and “Do you feel the ‘goldfish’ left anything out? If yes, what?” The T indicates that this activity closed “the pre-vacation leg of the unit” (25, L).	

Comparison:[[In both portraits, we see the T provide a variety of ELA experiences for these Ss. Reading, writing, speaking and listening happen on most days represented in the pf and in the CS.

Configurations are varied—small group, whole class, and pairs, in both self-selected and T-selected groupings.

We also see lessons that build on one another toward completion of a final composition. In the pf, those compositions are a character diary and a poem, and in the CS the final composition is a revision for on-line publication of a S-selected favorite piece of writing. While the goal in each case was to create a product, the lessons focused on other important skills. Through handouts in the pf and a Power Point presentation in the CS, we see the T scaffold students’ learning. She asks questions which aim toward having students achieve particular goals, but the questions themselves do not lead students to particular answers. In both portraits, we see the T guide her students in developing their own critical thinking skills. And, in both the pf and the CS, we see the T attend to development of students’ metacognitive skills.

Also, in both portraits we see established routines that Ss seem to respond well to and which seem to align with T goals. In the written pf text, the T tells us how students respond to particular activities, and in the video, we see that what she says is true. The CS provides additional evidence of the same. Both portraits would likely lead to the same evaluation of this T.]]

Classroom Interaction

Based on video	*Based on fieldnotes*
In the video conferencing session, we see the T seated at a table with four students. The procedure is that each S reads his or her poem aloud to the group, and group members provide feedback. The T facilitates by repeating and clarifying S responses after each one speaks, and by deciding	[Whole-class interaction in both the pf and the non-pf class is described by the CSW as representing the same pattern. The following example is one of several that illustrate that the T poses a question and Ss respond one after another with a variety of ideas.] < After Ss have read a sample the T has provided as part of her Power Point lesson, she poses the

when to move on to the next writer. Students are clearly practiced in this type of session. They make comments such as, “When you say ‘things,’ you could go deeper,” and “You know how you said it’s confusing? When I went over mine, I thought the same thing….” A student also defends his poem which two classmates say is “too deep” with, “That’s what I was aiming for” and explaining, “I’m trying to stay away from the word ‘like.’” The T makes comments such as, “What I’m hearing you say is that we have a good poem here, we just need …” and she ends the session by telling the students “I’m very pleased” with the way in which they conferenced.	question, “Does the entry show deeper thinking?” [A quality they had already determined was necessary for a good journal entry.] The CSW reports: <Mark attempts to give an example from the sample selection but his answer is not met with further agreement. Margaret states that the piece is boring and the teacher does admit that she feels it’s dull as well. Alex claims that the piece just tells about a fight with a friend, yet claims the reader really doesn’t know what the writer is talking about because it’s lacking in feelings. Josh says that it’s missing elaboration and just doesn’t make sense. Sabrina openly states, “It is a diary entry and simply that” and the teacher agrees that the piece is lacking in deeper thinking and asks if it paints a picture. The class offers a resounding NO and Stella chimes in that it just kept listing things. George says that the writer got mad and stayed mad without giving details. Another student states that the piece has no elaboration nor does it have sensory details. Brittany says she couldn’t feel empathy with the entry and just read it, not felt it. The teacher then asks the students how many of them have ever fought with friends and shared similar experiences> and the class moves on to another topic (CSW, 15).

Comparison: In both the pf and the CS, we observe similar classroom routines. We see the first 10 minutes of each class period spent on SSR in the pf, notebook writing in the CS. In both portraits, we see the T begin class sessions by asking a S volunteer to recap the previous day’s learning, then proceeding with the day’s activities. We witness a variety of classroom configurations. A horseshoe arrangement of desks is seen on the pf video and is described by the CSW. We observe whole class discussion, peer cfs about writing, small group activities, T lecture, presentations, and preparation of written texts for publication. Class sessions frequently close with the assignment of homework or preview of the following day’s activities.

While on the video the classroom may be seen as quite controlled in terms of time management and change from one activity to another by teacher command, it is clear in the *content* of the talk that students are thinking and learning. Their questions and comments to peers in writing conferences and in whole-group discussions indicate that they have learned to talk with one another about ideas, to give one another feedback about writing, and to

question and compliment their classmates. These are observable in the pf, and are affirmed by the CSWs observations of classroom dialogue among Ss and T.

Teacher's reflection on classroom interaction	
Based on commentary In order to "keep the noise to a minimum," the T reports that for the videotaping, she chose to work with one cooperative group while other Ss worked individually. When videotaping was finished, students were "released to work in peer groups of their own" (48). The T says that "each member of the group eagerly offered his or her perceptions on the shared work." She also reports that she questioned one Ss understanding of the assignment, but otherwise felt that Ss "were fully in command of the assignment." She offered her perspective to and moderated the discussion, but "made a conscious attempt not to influence student critiques" even when she felt that a S "was being a bit too literal in his images" (48). Ss have been in revision groups before, but the T notes that this is the first time they have written and revised poetry this year. Student Y is one of what the T terms "gray children," a student she says she needs to "watch particularly for" and "make a concerted effort to draw into class discussions and activities," as they readily "fade into the background" among more assertive Ss, and "do not cry out for the attention the lower-ability Ss require" (31-2). *[[What I see on the video jibes with the T's description of it. The writing session does feel*	*Based on interview notes* Reflecting upon a day's lesson, the T says she was surprised by 1st period's response: "They aren't generally that participatory" and she <admitted to perhaps selling them short as a group. She felt that perhaps the presence of a stranger motivated them on. In addition, she felt that the lesson ran its course the way she felt it would and she was disappointed that they didn't remember all of the qualities of good writing gone over since the beginning of the year. However, she recognized that the only quality period one struggled with was defining honest writing. She quoted, "One out of four isn't bad."> The T <felt that the students in period two demonstrated typical class behaviors by showing high levels of participation, enthusiasm for the lesson content, expressing that this is what she expects from period two. She further admits to being thrilled with the level of response from period one, yet states that "they have their on days and off days and you can jiggle the switch but the light doesn't go on." She can't figure out why this happens at times stating, "that it can't be the kids because they're not low level kids. They're average kids." It's a different personality and she feels these children don't know her as well as the other classes know her as they only experience her one period a day versus her other two classes which both have her for two periods in a day. She describes period one as "moving differently, flowing. They're far more vested in each other than me as a teacher"> (CSW, 31). When the CSW asked the T <to describe

somewhat rushed. The T did try to "facilitate" the writing conference group discussion, but she acted somewhat more as "guide" for the fishbowl activity, seemingly to push the Ss thinking. She says that the first group was nervous, and two of the members we see on camera do seem very nervous.]]	her interaction with a boy from period two who has a gift for analyzing detail and working with words. She comments, "He has written poetry that would stop your heart but he holds himself to a standard that is three times higher than any other child. He worries about things and he is able to grasp things that never even occur to the other kids. It works against him sometimes and needs a lot of reassurance and validation. Sometimes he just thinks too much and at times needs to have limits placed on him because of his drive and tremendous efforts." Teacher encourages him "to relax rather than get an ulcer before the age of twenty."> (CSW, 33).
Comparison: In the pf, we have info. about how the T views individual Ss as she writes about the 5 writing samples included in her pf. In this CS, the T also talks a great deal in interviews about individual Ss. In both cases, she talks both about who the child is as a person (e.g. personality, affect) and who the child is as a student (e.g. academic strengths and weaknesses). In both portraits we also see a range; the T provides information both about students who are the most successful and about students who struggle the most in her classes. In both portraits we also see this T's use of terms like "abstract randoms," and "concrete learners," and she describes herself as "concrete random." She appears to shape activities with these "types" in mind, as she does talk about "types" of Ss who will succeed or struggle with particular activities. She also describes individual Ss in terms of "high ability," "lower ability," and "average." There is no indication, however, that she holds Ss to particular standards based on which of these she believes the S to be. And, while the Ts characterizations of each class' personality differs, the CSW reports that "the teacher has the same basic instructional strategies for all of her classes" (CSW, FN 9).	
T's assessment of student work	
Based on comments written on students' papers On Student X's paper, the T circles 3 spelling errors and one missing comma. At the bottom of the page she writes 4 comments: Student X, much of this is wording	*Based on fieldnotes and interview notes* [[While we see no evaluated S writing in the CS, the T reviewed with Ss the qualities of good writing before (and during) their reading, selecting, conferencing, and revising processes for the task of polishing a journal entry for publication.]]

taken almost directly from the book; that's not the best tactic here. Use your own words. Not clear whose POV you're taking in this entry . . . What about this character's thoughts and feelings? this is mainly summary. Reflect on the events—don't' just list them! *Based on T's commentary* In reviewing the Ss initial character diary entries, the T reveals, "I feel I may have overestimated the ability of many Ss to take another's point of view. The work of Student X in particular reminds me that they are still young, and I wonder if some Ss have not yet matured enough to see beyond themselves, to look at things from another's perspective. Student X's entry shows that he is taking the book very much at face value; he either cannot see (or does not wish to take the effort to see) the character traits of the characters conveyed in their words or actions" (33).	The CSW writes that the T <starts with discussing the ingredients of a good notebook entry: deeper thinking, painting a picture with words, the value of coming up with one's own ideas, not writing diary type words and striving for that full page of writing> (FN 45). [[Two artifacts appended to the pf provide the more specific detail.]] A slide from the T's Power Point presentations asks: Does the entry show "deeper thinking?" Does the writer "paint a picture with words?" Is the writing "honest writing?" Is the topic unique, unusual, or particularly meaningful? The Nomination Ballot students use to evaluate their own and a classmates' work lists the following criteria for judgment in addition to the above: Entry is clear and focused Wonderful use of "show, don't tell." Captivating written "voice."
Comparison: It seems that even though we do not see the Ts actual feedback on Ss work in the CS, the nature of her conversation with them and the content of the slides and the evaluation sheets Ss are to use with peers make us confident that she will actually use these standards to evaluate the final products. So while we know more detailed information, and see the actual follow-through to final draft only in the pf, there is nothing in the two portraits that would likely cause this T to be evaluated differently in regard to what she thinks is important in evaluating S writing.]]	
Teacher's reflection on her teaching	
The T reports that when she begins this unit another time, she	<In regards to long range goals, the T reports that this is the first time she has

will spend more time on particular aspects of studying of the novel that she felt Ss needed more time with. For example, she says that next time she will “spend one day focusing only on the aspects of fantasy before launching into the hero.” She combined the two in this lesson, and reflects: “it would certainly have benefited from additional time for student work and discussion” (L, 7). The T notes that a vocabulary glossary or a vocabulary activity “in preparation for the reading” might be useful to Ss, as they asked about several words as the T read chapter 1 of the novel aloud. The T lists “prodigious,” frenzied,” and “exclusive” as examples (L, 10-11). The T also says she “may have overestimated the ability of many students to take another’s point of view.” Her conclusion is, “In the future, I should probably ‘test drive’ the character diary on a short story before applying it to a longer piece, so as to better assess the abilities of the class to step into another’s point of view” (33).	tried this unit, yet feels she might do the same unit again next year and states that she will give the introductory portion in one day and the nomination ballot for homework. She also believes that the selections regarding best pieces will be done in class ‘to eliminate the consequences of Ss who come to class unprepared’> (CSW, 33). < During the three day observation period, this interviewer was able to witness her creating curriculum from day to day based on her monitoring and adjusting for student learning first-hand. Her Power Point demonstration on revision, for instance, was created in response to her feeling that the students needed it to complete the task at hand> (CSW 29).

GQ 2.1 Describe the ways in which the teacher creates a learning environment that provides all Ss with opportunities to develop as readers, writers, and thinkers:

[[In the pf, the T describes “types” of S learners (e.g. concrete, random) in her logs, her other written text, and her descriptions of those whose writing samples she includes. She speaks in both the pf and the CS about accommodations for special needs Ss (e.g. special education), and the CSW observes the T’s accommodations for absent Ss.

In both the pf and the CS, we observe similar classroom routines. We see the first 10 minutes of each class period spent on SSR in the pf, notebook writing in the CS. In both portraits, we see the T begin class sessions by asking a S volunteer to recap the previous day’s learning, then proceeding with the day’s

activities. We witness a variety of classroom configurations. A horseshoe arrangement of desks is seen on the pf video and is described by the CSW. We observe whole class discussion, peer cfs about writing, small group activities, T lecture, presentations, and preparation of written texts for publication. Class sessions frequently close with the assignment of homework or preview of the following day's activities.

While on the video the classroom may be seen as quite controlled in terms of time management and change from one activity to another by teacher command, it is clear in the *content* of the talk that students are thinking and learning. Their questions and comments to peers in writing conferences and in whole-group discussions indicate that they have learned to talk with one another about ideas, to give one another feedback about writing, and to question and compliment their classmates. These are observable in the pf, and are affirmed by the CSWs observations of classroom dialogue among Ss and T.

The T remarked in the pf text and in the CS interviews that some of the activities she planned may be (or may have been) too difficult for some of the Ss. When she anticipated that in advance, she attempted to shape instruction accordingly. When she realized Ss difficulties during or after class, she altered her plans for the next day (e.g. created another handout, attended to another aspect of the writing process in her Power Point presentation), or she indicated how she will alter instruction in the future. In both portraits, the T creates activities that challenge Ss, and her questions challenge their thinking on handouts and in discussion activities. In addition, we see the T make minor alterations in a particular lesson from one class to another, but she indicates and the CSW observes that she works with the same lesson plans, assignments, and "talk" as she guides S learning. In each individual class, however, the T asks questions and pushes discussion based on Ss contributions, thus responds in flexible ways to their learning as a class and as individuals within that class.

For this T, the CS reinforces what we learn in the pf. Either portrait alone would have given a substantial amount of information in regard to this Guiding Question, and neither contains information that would likely lead us to evaluate this T differently.]]

GQ 4.1 Describe how the teacher addresses student learning in reflection.

GQ 4.2 Describe how the teacher uses that reflection to inform practice.

[[This T, in both the pf and the CS, addresses issues of S learning in a variety of ways, including addressing learning in terms of the class as a whole and the individuals within the class. She reflects on students in general (age/developmentally) as well as on what she has learned about her own students this year.

In both the pf and the CS, we see the T reflect on student learning in terms of the nature of their contributions to discussion. In the pf, she reflects on S

learning in terms of their “command” of assignments and of specific tasks, and in terms of the depth of reflection and the insightfulness shown in their writing. In the CS, she reflects on Ss abilities, their strengths and weaknesses, and the behavior they exhibit during class sessions. So, while we gather some similar and some different information about how the T reflects on S learning in each of these portraits, we can see in both that she draws upon a range of information that is both accurate and important in informing her instruction. None of the differences between the evidence presented in the pf and that presented in the CS is likely to be the cause of a different evaluation for this T.]]

The World Wide Web address for the *Education Policy Analysis Archives* is
epaa.asu.edu

Editor: Gene V Glass, Arizona State University

Production Assistant: Chris Murrell, Arizona State University

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. The Commentary Editor is Casey D. Cobb: casey.cobb@unh.edu.

EPAA Editorial Board

Brasil
Email: gfrigotto@globo.com

- Vanilda Paiva
  Email:vppaiva@terra.com.br
- Lilian do Valle
  Universidade Estadual do Rio de Janeiro, Brasil
  Email: lvalle@infolink.com.br
- Romualdo Portella do Oliveira
  Universidade de São Paulo, Brasil
  Email: romualdo@usp.br
- Roberto Leher
  Universidade Estadual do Rio de Janeiro, Brasil
  Email: rleher@uol.com.br
- Dalila Andrade de Oliveira
  Universidade Federal de Minas Gerais, Belo Horizonte, Brasil
  Email: dalila@fae.ufmg.br
- Nilma Limo Gomes
  Universidade Federal de Minas Gerais, Belo Horizonte
  Email: nilmagomes@uol.com.br
- Iolanda de Oliveira
  Faculdade de Educação da Universidade Federal Fluminense, Brasil
  Email: iolanda.eustaquio@globo.com
- Walter Kohan
  Universidade Estadual do Rio de Janeiro, Brasil
  Email: walterko@uol.com.br
- María Beatriz Luce (1998—2003)
  Universidad Federal de Rio Grande do Sul-UFRGS
- Simon Schwartzman (1998—2003)
  American Institutes for Resesarch–Brazil

Canadá

- Daniel Schugurensky
  Ontario Institute for Studies in Education, University of Toronto, Canada
  Email: dschugurensky@oise.utoronto.ca

Chile

- Claudio Almonacid Avila
  Universidad Metropolitana de Ciencias de la Educación, Chile
  Email: caa@rdc.cl
- María Loreto Egaña
  Programa Interdisciplinario de Investigación en Educación (PIIE), Chile
  Email: legana@academia.cl

España

- José Gimeno Sacristán
  Catedratico en el Departamento de Didáctica y Organización Escolar de la Universidad de Valencia, España
  Email: Jose.Gimeno@uv.es
- Mariano Fernández Enguita
  Catedrático de Sociología en la Universidad de Salamanca. España

Email: enguita@usal.es
- Miguel Pereira
Catedratico Universidad de Granada, España
Email: mpereyra@aulae.es
- Jurjo Torres Santomé
Universidad de A Coruña
Email: jurjo@udc.es
- Angel Ignacio Pérez Gómez
Universidad de Málaga
Email: aiperez@uma.es
- J. Félix Angulo Rasco (1998—2003)
Universidad de Cádiz
- José Contreras Domingo (1998—2003)
Universitat de Barcelona

México

- Hugo Aboites
Universidad Autónoma Metropolitana-Xochimilco, México
Email: aavh4435@cueyatl.uam.mx
- Susan Street
Centro de Investigaciones y Estudios Superiores en Antropologia Social Occidente, Guadalajara, México
Email: slsn@mail.udg.mx
- Adrián Acosta
Universidad de Guadalajara
Email: adrianacosta@compuserve.com
- Teresa Bracho
Centro de Investigación y Docencia Económica-CIDE
Email: bracho dis1.cide.mx
- Alejandro Canales
Universidad Nacional Autónoma de México
Email: canalesa@servidor.unam.mx
- Rollin Kent
Universidad Autónoma de Puebla. Puebla, México
Email: rkent@puebla.megared.net.mx
- Javier Mendoza Rojas (1998—2003)
Universidad Nacional Autónoma de México
- Humberto Muñoz García (1998—2003)
Universidad Nacional Autónoma de México

Perú

- Sigfredo Chiroque
Instituto de Pedagogía Popular, Perú
Email: pedagogia@chavin.rcp.net.pe
- Grover Pango
Coordinador General del Foro Latinoamericano de Políticas Educativas, Perú
Email: grover-eduforo@terra.com.pe

Portugal

EPAA is published by the Education Policy Studies Laboratory, Arizona State University