EDUCATION POLICY ANALYSIS ARCHIVES

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

Volume 12 Number 33 **July 20, 2004** **ISSN 1068-2341**

Reasonable Decisions in Portfolio Assessment: Evaluating Complex Evidence of Teaching

Aaron Schutz
University of Wisconsin-Milwaukee

Pamela A. Moss
University of Michigan

Citation: Schutz, A., Moss, P.A., (2004, July 20). Reasonable decisions in portfolio assessment: Evaluating complex evidence of teaching, *Education Policy Analysis Archives, 12*(33). Retrieved [Date] from http://epaa.asu.edu/epaa/v12n33/.

Abstract

A central dilemma of portfolio assessment is that as the richness of the data available to readers increases, so do the challenges involved in ensuring acceptable reliability among readers. Drawing on empirical and theoretical work in discourse analysis, ethnomethodology, and other fields, we argue that this dilemma results, in part, from the fact that readers cannot avoid forming the data of a portfolio into a pattern—a coherent "story" or "stories"—in order to evaluate it. Our article presents case studies of readers independently evaluating the same portfolios. We show that even readers who hold a shared vision of effective teaching and who cite much the same evidence can, nonetheless, develop significantly different "stories." Our analysis illustrates that some portfolios are more ambiguous than others and are thus more likely to result in such divergent readings. We argue

that more fine grained understandings of portfolio ambiguities and disagreements between readers over "stories" can help us respond to the challenges posed by the rich data of portfolio assessments. (Note 1)

Introduction

Imagine that you are sitting in the back of a classroom. Every passing moment is rich with events. Students fidget, make marks on paper, and whisper side comments to each other; the teacher asks questions, draws on the board, spreads her arms to indicate the size of an elephant; the principal makes an announcement over the intercom; a book falls from a desk; a passing cloud dims the sunlight coming through the windows.

Then imagine it is your job to assess the teacher's performance in this classroom. You immediately face a wide range of challenges and dilemmas. For your interpretation to be well-warranted, you must attend to a variety of relevant evidence. But interpretations like these always involve some level of selective abstraction and pattern-making. Some aspects must be foregrounded from the near infinite range of the noticeable, and this means other aspects will fade into the background. Further, as this paper will show, even when you decide what evidence matters, you must draw this together into a comprehensible pattern. And no matter how much evidence you collect, it represents only a small sample of the information that could be engaged with, always reflecting a particular understanding of what is important and what is not. In this article, we seek to illuminate some of thesechallenges and dilemmas involved in moving from evidence like this to well-warranted interpretations in consequential assessments of teaching.

The practice of assessment in American schools has been largely guided by the discipline of educational measurement. Over time, measurement scholars have developed a shared set of strategies for drawing and warranting assessment-basedinterpretations--for “reasoning from evidence to inference” (Mislevy, Almond, and Steinberg, 2003). As Mislevy and colleagues noted, the general approach reflected in these strategies has been driven, in part,by the need to make conclusions about large numbers of individuals. By standardizing their approach, by collecting the same data from each person under the same conditions and using the same criteria and procedures to evaluate it, assessment developers can build a common argument for validity rather than having to reinvent the argument anew for each case. When an assessment system is operational, then, the set of possible scores (and the intended interpretations associated with them) is essentially predetermined: the goal is to use the available evidence to determine the most appropriate or likely score/interpretation for each individual. Importantly, however, assessment scholars understand that the common argument developed for any particular test will not necessarily hold for all individuals. Thus Mislevy and colleagues noted that assessors actually have dual responsibilities in their efforts to construct arguments for validity. Not only must they establish "the credentials of the evidence in the common argument," they must also detect "individuals for whom the common argument does not hold” (p. 15).

In this paper we focus on one crucial component of a common validity argument for performance assessments: the role of readers (raters, assessors, or judges) in interpreting/evaluating the available evidence about an individual. In particular, we focus on the role of readers in portfolio assessments that contribute to consequential decisions about individual teachers. (Note 2)

Most assessment systems seek to maximize the consistency with which readers are applying the scoring rubric and to minimize threats to validity that might be introduced as readers score. Developers work to ensure readers attend only to evidence relevant to the construct of interest (minimizing "construct irrelevant variance") and capture all relevant evidence andcriteria (reducing “construct under-representation”), developing ongoing monitoring programs to ensure the system is actually functioning as intended. [Messick (1989)providedextended descriptions of these general threats to validity, which AERA, APA, NCME (1999)echoed.] Consistency is typically promoted through training and monitoring procedures that ensure,to the extent possible,that readers are using the same appropriate criteria to attend to the same relevant features of the available evidence (see, e.g., Engelhard, 1994, 2001; Myford and Engelhard, 2001; Myford and Mislevy, 1995; Wilson and Case, 1997).

As Kane, Crooks, and Cohen (1999) noted, however, “the more complex and open-ended the task, the more difficult it becomes to anticipate the range of possible responses and to develop fair, explicit scoring criteria that can be applied to all responses” (p. 9), and thus the more difficult it is to prepare readers able to consistently apply their rubric. Teaching portfolios of the kind we discuss contain some of the most "open-ended" data encountered in large scale assessments, including contextualizing commentary, videotapes of teaching, instructional artifacts, samples of student work, and the teacher’s reflective commentary. A central dilemma of portfolio assessment is that as the richness of the data available to readers increases,so do the challenges involved in ensuring acceptable reliability between readers. The very thing that would seem likely to improve the validity of an assessment--more information--also appears to threaten its validity.

We engage with this dilemma, here, by seeking to better understand the kinds of interpretive challenges that portfolios seem to raise. These are often illuminated in disagreements between readers. However, the sorts of ambiguities that underlie disagreements can also be present in portfolios even when particular readers agree on a score, especially when the scoring system constrains or reduces the complexity of the judgments that readers are required to make. We argue thatmore fine-grained understanding of such disagreements and ambiguities can help us improve our general arguments for validity, identify cases where the argument does not fit, and pointus to new strategies for ethically taking account of such issues in the context of large scale assessment.

While we focus in this paper on portfolio assessments of teachers, the issues we raise are just as relevant for other forms of assessment, including essay exams and multiple choice tests. In multiple choice tests, for example, the task of making "sense" of the data collected is given over to a key which entails assumptions about what students mean when they answer each question and about how the combined picture provided by all of a student's answers together

should be understood (e.g., Hill and Larsen, 2000). We examine portfolios, then, not because they are inherently more challenging or problematic from an interpretive point of view than simpler forms of assessment, but because the burdens they place on readers tend to illuminate challenges otherwise obscured by less open-ended assessment contexts. Thus, the crucial difference between a multiple choice test and a teaching portfolio is not the complexity of the *performance* they represent--the reality they are attempting to describe---but the complexity and richness of the *data* available to readers about that performance. Portfolio assessments like these allow us to see what happens when much of the complexity of a performance has not been eliminated by the assessment instrument itself.

Our data are drawn from field tests of a portfolio assessment developed by the Interstate New Teacher Assessment and Support Consortium (INTASC) and adapted for use in Connecticut. Building on the pioneering work of the National Board for Professional Teaching Standards, INTASC is developing subject specific standards and portfolio assessmentsto help participating states support the professional development of beginning teachers and make licensure decisions. Of the 10 INTASC states that participated in the development of the portfolio assessment, only Connecticut is currently using portfolios to inform licensure decisions. The portfolios under consideration were prepared by beginning English/language arts teachers as part of a special field test in Connecticut. In this field test, readers (experienced teachers in the subject area) evaluatedportfolios in pairs, seeking consensus on a final score. We present two case studies of three pairs of readers each reading the same portfolio. The three pairs of readers disagreed quite widely on the first portfolio (scores of 1, 3, and 4 on a scale of 4, with 2 considered a passing score) but agreed (on a score of 1) on the second portfolio.

Our analysis indicates that many of the disagreements between reader pairs that might initially seem to be about “evidence" (are readers attending to the relevant features of the performance?) or "values" (are they applying the same criteria?) actually represent disagreements over what we term the “story” of the portfolio. We use the term "story," drawn from narrative theory and studies of the interaction of narrative and human understanding (e.g., Bruner, 1986; Davies and Harre, 1990; Kroeber, 1992), to indicate the patterns people develop to make coherent sense of situations and texts. In many cases,although readers appeared to agree on the what was "there" in the portfolio and largely concurred on their vision of what counted as effective teaching, they nonetheless constructed different "stories" that fit much the same data into very different patterns. Sometimes, for example, a divergence appeared to involve one pair foregrounding an aspect of the performance that the other pair tended to downplay. For example, in the Civil Rights Movement (CRM) portfolio, discussed below, while one pair (Robert and Sandra) acknowledged that most of the dialogue in the Response to Literature (RTL) video involved simple “question and answer,” they stressed particular examples where the teacher went beyond this. Another pair, Charlene and Iris, focused on the opposite characteristics. They stressed the “question and answer” aspect of the dialogue, even though they noted as exceptions aspects the first pair had emphasized. Here, the readers appear to have viewed much the same “evidence” and to have held similar “values” about what counted as better and

worse teaching. Yet they still constructed fundamentally different pictures of what was going on in the classroom. And each pair made convincing arguments for their points of view.

Most operational assessment systems depend upon evidence of discrepant readings to illuminate problems like these for further review. And yet most large scale assessment systems also work hard to minimize the number of discrepant readings--to improve the consistency with which readers are applying scoring criteria. When inadequate levels of interreader reliability are found (as is more frequent with portfolio assessments) the advice for improving reliability typically includes disaggregating the portfolio into components that can be separately scored, having different readers score each component (Nystrand et al., 1993; Klein et al., 1995; Swanson et al., 1995) and "introducing separate scales to disentangle confounding aspects of performance" (Myford and Mislevy, 1995, p. 55). However, these are all practices that are more likely to gloss over rather than illuminate the sort of ambiguities in evidence we will demonstrate. Thus, our concern is not only that routine practices for building a reliable system fail to illuminate ambiguous cases, but that advice for improving the system to make it more reliable is likely to make the problem even harder to detect.

In the next section we provide a theoretical foundation for exploring disagreement over "stories," accompanied by empirical evidence from the limited set of studies that have examined actual reader processes in the context of large scale performance assessments. Then we turn to two case studies of multiple pairs of readers independently reading the same two portfolios. In our concluding comments, we speculate about the ways large scale assessment systems might feasibly and ethically cope with the sorts of problems we have raised.

Disagreements over "Stories:" A Theoretical Discussion

Our conception of disagreements over what we call "stories" draws on assumptions about how human beings make sense of the world from a diverse range of fields, including analytic philosophy (Grice in Davies, 2000), narrative theory in literary studies (Kroeber, 1992), research on reading processes (Ruddell and Unrau, 1994; Smagorinsky, 2001), historiography (Mink, 1987), linguistics (Gee, 1990), psychology (Bruner, 1986,1990; Gibbs, 1993), hermeneutics (Gadamer, 1975, 1987)and ethnomethodology and conversation analysis in sociology (Arminen, 1999; Garfinkel, 1967,2002; Schegloff, 1999). Despite differences, these scholars have converged, often somewhat independently, on a description of human understanding that involves the *active* construction of coherent meaning in our everyday interactions in the world. In this section, we explore aspects of the arguments that lie behind this theoretical vision and illustrate these arguments with evidence from the relatively small body of literature that examines, inductively, readers' processes in large scale assessment (e.g., Heller, Sheingold, and Myford, 1998; Moss, Schutz, and Collins, 1998; Myford and Mislevy, 1995). We primarily focus on ethnomethodology, especially the efforts of Harold Garfinkel (1967, 2002), because, along with other work on narrative, its empirical findings anticipate with remarkable accuracy the documented inclinations and practices of readers when they encounter complex performancesin large scale assessments.

Furthermore, it provides a fruitful theoretical resource for illuminating particular problems with existing assessment systems and for designing assessment systems that better accommodate complex performance data.

Garfinkel and colleagues (1967,2002) arguedthat theprocess of sense-making, of imputing meaning to data, intention to actions, and the like, is a constant and permanent aspect of our efforts to orient ourselves in the world. We do this through what he called the “documentary method.” The documentary method “consists of treating an actual appearance of something”--for example, a view of an object, a statement in an ongoing dialogue, an action, a piece of text from a portfolio, or the portfolio itself—“as ‘pointing to,’ as ‘standing on behalf of’ a presupposed underlying pattern” (p. 78). In other words, we make sense of what something *is* or *means* by referring to an assumed underlying pattern of which it is a part. Conversation analysts (see Heritage, 1984), for example, have often shownthat we are able to understand quite cryptic statements made by our conversation partners in everyday dialogue only because we make assumptions about what our partners are talking “about.” As we are leaving a movie, when my friend says “I didn’t like it,” I need to know enough about her biography, the context, and our past history of conversations to know whether she is referring to the popcorn, the movie, her job, or any number of an infinite number of possibilities. What is interesting to these discourse analysts is something we generally take for granted: that despite the seemingly complex interpretive work needed to make sense of such ordinary interactions, people generally understand each other and their environments quite well.

Garfinkel (1967,2002) arguedthat all statements, no matter how detailed we try to make them, are essentially “*indexical*.” In other words, like my friend’s statement about what she didn’t like, they point to something never entirely contained within the statement itself. And this is not just a problem with pronouns like “it;” this is a pervasive feature of human interaction. In experiments, Garfinkel (1967,2002) found that no matter how hard one tries to specify exactly what any statement “means,” there is always some ambiguity left. His students, for example, when trying to answer a question like “What do you mean you ‘had a flat tire’?” would invariably throw up their hands at some point and declare that the meaning in a particular example was something “anyone can see.” Understanding, he argued, always involves reference to assumptions not contained explicitly in a dialogue or a text.

Importantly, Garfinkel (1967)found that this process of sense-making is *recursive*. Not only is “the underlying pattern derived from its individual documentary evidences,” but “the individual documentary evidences, in their turn, are interpreted on the basis of ‘what is known’ about the underlying pattern” (1967, p. 39). Or, as he notes elsewhere, “you can speak either of seeing the detail-in-the-pattern or the pattern-*in*-the-detail” (2002, p. 203). This is analogous to Gadamer’s (1975, 1987) “hermeneutic circle” in which we discover what a text means by analyzing the parts, and give meaning to the parts by understanding the whole. (Note 3) Consider the following artificial example (taken from Wittgenstein, in Heritage, 1984, p. 87). What do you see?

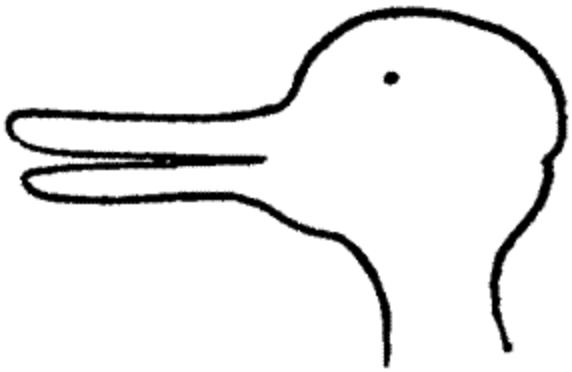

Figure 1: Line Drawing

(Wittgenstein, in Heritage, 1984, p. 87)

In this very simple drawing, two fundamentally different conclusions seem possible. Either the picture is of a duck, or it is of a rabbit. And each story about the meaning of the picture involves a gestalt switch in the meaning of all of the data in the picture. In the case of the duck, for example, the indent on the right is merely a dent in its head. While we all may "see" it, it probably isn't important enough to be given much attention. The duck would still be a duck without it. On the other hand, when we see a rabbit, the indent becomes the rabbit's mouth. Suddenly it is one of the most important features in the picture. In fact, if you put your finger over the indent, suddenly a duck is the only reasonable interpretation. The only reason you still see the possibility that there could be a rabbit is because you can't forget about the indent. If you show the picture to someone else with the indent covered, they will almost invariably tell you that, as "anyone can see," it is a duck.

Actually, you can try a third approach and attempt to see the picture as simply a collection of curved lines, trying not to see it either as a rabbit or a duck. You will probably find this is a difficult task—you cannot easily take "time out" from interpretation, from seeing "what everyone can see," even when you try. *Not* interpreting feels like the *active* stance since the largely unseen process of sense-making is our default mode of interaction with the world. We'll revisit the issue of ambiguous evidence--of the potential to support multiple stories--that this example raises shortly.

Returning to the world of human interaction, our present analysis and previous work shows that such recursive efforts to "make" sense occur continually in readers' efforts to understand portfolios, even though, like all of us, they generally seem not to notice the constant, moment-to-moment judgments they are making (Moss, Schutz, and Collins, 1998; see Goodwin and Goodwin, 1992). These natural inclinations to make sense have also been illustrated repeatedly in the small body of literature that looks inductively at how raters reason in evaluating open-ended assessments, both single performances (Freedman and Calfee, 1983; Wolfe, 1997) and portfolios (Heller et al., 1998; Moss et al., 1998; Myford and Mislevy, 1995). All of these studies found readers working recursively between interpretations/evaluations and chunks of text. As described by Heller and colleagues (1998), Freedman and Calfee found that

bits of text are judged as they are comprehended and that “the evaluation of one section of text may change as one comprehends a subsequent section of text” (p. 4). They noted that Wolfe (1997) “confirmed that some raters exhibit iterative processing and intermingle reading, evaluating, and articulating rating processes when judging essays” (p. 4). Heller and colleagues (1998), listening to think alouds while readers rated student portfolios, found readers drawing on whatever they could find that was “remotely useful” to fill in the gaps in their understanding (p. 17).

For example, raters referred to the dates on which pieces of student work were created when deciding how to interpret qualities of performance across versions of a piece that had been revised . . . or across different pieces in a portfolio over time (p. 17)

They noted that “repeated exposure to portfolios from the same class was also useful. For instance, it allowed them to draw inferences about how the teacher has structured the assignments and “what the child has brought to the work” (p. 18; see also Moss et al., 1992).

In fact, ethnomethodologists have shown that it is very difficult to prevent people from engaging in this ongoing process of sense-making/story construction. For example, in a range of experiments Garfinkel (1967, 2002) attempted to “breach” participants’ understanding of the world. In one famous case, students met and talked with a counselor about an issue that was important to them. The counselor then left for a separate room and students asked a series of questions of the counselor through an intercom, receiving either “yes” or “no” answers. After each question, students shut off the intercom and talked into a tape recorder about their understanding of the answer. What is interesting is that the answers were given in a completely random order. Yet in nearly every case the students had little difficulty making sense of them. They noted that they “saw in a glance” “what the advisor had in mind.” They were able to manage incongruous answers by “imputing knowledge and intent to the advisor,” although sometimes they had to wait to understand the meaning of a particular answer. (Note 4) And each new answer might change the sense of earlier answers and even the sense of the student’s own questions. Thus “the sense of the problem [a student was focusing on] was progressively accommodated to each . . . answer, while [each] answer motivated progressively fresh aspects of [the student’s] underlying problem” (1967, p. 90; see Arminen, 1999).

A few key lessons can be taken from Garfinkel’s (1967) counseling experiment. First, as later work in conversation analysis showed in detail (Arminen, 1999), understanding in such an exchange is built up over time as people seek to build a coherent story. Sometimes the meaning of a statement or of a piece of data changes in response to new information, and sometimes the meaning of a particular piece of data remains vague until further information is forthcoming. In conversation, for example, “every next conversational move renews our understanding of the prior move, so that each turn at talk . . . recreates the context anew” (Arminen, 1999, p. 41). Once a particular pattern has been identified, subsequent pieces of data will be interpreted in the light of this pattern. Even if the subsequent data disconfirms the initial pattern, it is read to

some extent in response to the pattern it disconfirms. Once any descriptive work has been done, then, interpreters are moved from total uncertainty to the progressive development of a few anchor points from which to work. Thus, while readers, for instance,change their views as they read, and reinterpret previous evidence in the light of subsequent evidence, the *sequence* of their interpretations has a crucial effect on their subsequent readings (Arminen, 1999).

Our previous analyses of reader processes have shown much the same thing, that the more a reader "sees" a particular pattern, the more she may be inclined to interpret further evidence as "pointing to" this pattern—even if, to an outside observer with access to multiple interpretations, the data might seem more equivocal. Further, while a reader may hold a few pieces of evidence in a state of uncertainty as to their meaning, like the students in the counseling example, it would seem difficult for them to hold very many in this state given the time pressures they are under and the amount of data they are required to examine. In fact, in an earlier study of this approach to portfolio assessment,we found that initial interpretations reached by a reader pair often "cascade" through their subsequent reading of a portfolio, influencing how they "see" future pieces of data (Moss, Schutz, and Collins, 1998). As we note below, however, and as this discussion indicates, such a "cascade" does not necessarily involve a problematic rush to "prejudgment" on the part of evaluators. As the theory we have cited suggests, readers cannot avoid interpreting parts in light of some understanding of the whole. Nonetheless, readers in our case studies often sought out evidence that might contradict the patterns they were building.

Second, students in Garfinkel's counseling experiment expended a great deal of interpretive energy to eliminate contradictions and to ascribe a coherent "sense" to all of the responses they received, something he saw in his later work as well (Garfinkel, 2002). Even when students were told the nature of the experiment, they had great difficulty treating the advisor's responses as "random." "Many subjects," Garfinkel (1967) noted, "saw the sense of the answer 'anyway'" (p. 91). As Heritage (1984) pointed out, "it is striking that, despite the ingenuity of many of the experimental procedures [Garfinkel and his colleagues engaged in, his goal of creating situations that were literally unintelligible], was rarely achieved" (p. 94).

Finally, students in the counseling experiment had difficulty *not* imputing *intention and motivation* to the counselor's statements. In fact, in additional experiments, when seemingly senseless actions were taken—like when Garfinkel (1967) replaced one white pawn with another white pawn while playing chess—participants immediately sought to understand the motivation behind the action within the goals of the game, even after it was explained to them that there was no such motivation. Readers of portfolios assessments have, similarly, sought to understand students' intent: Heller and colleagues (1998), listening to think alouds while readers rated student portfolios in different subject areas, found them trying to understand what the student had done or was asked to do by the teacher: 'Before I can evaluate the student's work, I have to know what the problem is that they're attacking.' (p. 16)

Along the same lines, Myford and Mislevy (1995), interviewing readers about

difficult-to-score portfolios in the Advance Placement Art program, found readers raising concerns about not knowing enough about the students' intent, especially in the case of uneven portfolios. For instance, they described readers expressing their frustration over not being able to talk to certain students directly about the work they had submitted. They felt that such conversations would help them answer crucial questions they had about the students' work, and would enable them to feel more confident about their judgments, particularly in cases in which they were vacillating between ratings. (29)

At this point, readers might expect us to make a broad argument about the impossibility of valid portfolio interpretation, something like the simple relativist argument sometimes made by some recent readers of postmodernism. However, such an argument would ignore a basic truth of human existence. Most of the time we do understand each other sufficiently to get along, and our interpretations of the world generally serve us well (Gee, 1990). It is, in fact, unreasonable to say to someone "what do you mean, 'a flat tire'?" Of course we know what they mean, at least enough to go on with the conversation. Understanding our world and other people is what we *do* moment to moment, sitting in chairs because they are chairs, deciding not to cross the street right now because it is too dangerous, replying to questions because they make sense. This is not to deny that we sometimes make mistakes, or are sometimes confused, but most of the time we can make what Garfinkel called"reasonable" judgments, which he definedas "outcomes of documentary work, decided under the circumstances of common sense situations of choice," where "common sense" in the case of portfolio evaluation involves the use of tools of judgment provided to readers through training (Garfinkel, 1967, p. 99).

Of course, there is always the possibility that an individual will come to a relatively idiosyncratic interpretation in any particular case. The distinction, here, between "reasonable" and "unreasonable" is a qualitative one. It relies on interpreters holding basically the same set of values or general "methods" for interpreting. Every person has a unique biography and cannot help but make "sense" of the world in their own unique way to some extent, regardless of how they are trained. Thus, perfect reliability of interpretation is never achievable. Whether a particular interpretation is "reasonable" cannot be proved, as "reasonable" is not a scientific but a social designation achieved in the same way as all other "documentary" efforts.

Sometimes however—as in the Duck/Rabbit example described above—we find it difficult to come to "reasonable" decisions that "everyone can see." In fact, we argue that in these specific circumstances, it is sometimes "reasonable" *not to be able to decide on a particular interpretation*. For example, Heritage (1984) discusseda group that was contracted by the government to decide whether particular cases of death were examples of suicide. Such an effort seems inherently fraught with uncertainty. People who commit suicide may not have particularly clear intentions, they may have motivations for hiding their true intentions, and a characteristic of some suicidal people is that they cannot communicate what is bothering them very well. They may even appear happier just prior to suicide. Heritage madereference to a particular example that is illuminating. In this case, a woman was found asphyxiated in her apartment with the gas on and towels pushed against all the windows and spaces under the

door. On first glance, this looks very much like a classic case of suicide. But then Heritage gave some extra information: it was extremely cold that day, and her neighbors and friends said she was always a very happy person. This seems to open up contradictory possibilities. On the one hand, the woman may have turned the gas on to commit suicide, pushing towels against any drafts to make sure the gas stayed in her apartment. On the other hand, the woman may have been cold that day, and when it didn't get warmer (perhaps because the pilot light had blown out) she pushed towels against any drafts to keep the heat in and accidentally asphyxiated herself.

What we have, here, is a case where at least two fundamentally different "stories" appear to make adequate sense of the same collection of data. And each story (or "pattern") gives fundamentally different meanings to the different pieces of evidence available to us. In the first case, the woman used the towels to kill herself, in the second case she used them to keep warm. In the first case she intentionally blew out the pilot light. In the second case, she didn't know it was out. Each story requires a fundamental gestalt shift in the meaning of all of the data. Of course, there may be other equally convincing stories that could be told about the event as well. And it is not clear that *more* data will make judgment easier. In the suicide case, for example, more data has already made it more difficult to decide whether the woman committed suicide—and it is quite possible that additional data will only muddy the waters more. (Note 5) As in the Duck/Rabbit example, sometimes different stories can "reasonably" fit a particular collection of data. And the Duck/Rabbit example shows how different explanatory frameworks can even alter what *counts* as a "piece" of data. In one, the indent is not worth mentioning, is not really a separate aspect of the duck, in the other it is a crucial and separately identifiable body part, although we "see" it in both cases.

In fact, reader process studies indicate that readers often face challenges of interpretation and story construction similar to those just noted, especially when faced with inconsistent or ambiguous evidence. In Myford and Mislevy's (1995) example of a portfolio assessment that included art work and commentary on that work, for example, disjunctions between the two data sources created interpretive problems. When insightful commentaries did not seem to reflect what could be seen in the actual work, or when weak commentaries accompanied a series of high quality interrelated pieces, readers struggled to develop a coherent interpretation of the portfolio. In our own work on INTASC portfolios, we have repeatedly encountered similar challenges, for example, when relating teachers' commentary to classroom videos (Moss et al., 1998, 2003). Heller et al.'s (1998) study presentedsimilar examples of readers struggling with discrepant information. Interestingly, Heller et al. notedthat when readers couldn't find evidence to resolve the situation, they resorted either to quantitative solutions (e.g. averaging scores from separate pieces) or made individual decisions about how to weigh criteria and evidence. These strategies appeared to serve as fall back solutions when the evidence did not allow them to develop a single coherent representation (or "story" in our terms).

We draw on this conception of "stories" in our two case studies, illustrating the kinds of interpretive problems that complex portfolio assessments present for readers.

Assessment Context and Case Study Methodology

The INTASC portfolio assessments discussed here are intended for teachers in their first, second, or third year of teaching. To guide the development and evaluation of such portfolios, 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 pointing toward 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, videotapes 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 multiple separate entries (see, e.g., ETS, 1998), 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 6) As developed by INTASC, the portfolio assessment was intended both for professional development and for informing decisions about licensure. As implemented in Connecticut in 2000, there were four possible levels to the overall decision: conditional, basic, proficient, and advanced, where basic constituted a pass. 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.

During training, readers (experienced teachers in the relevant subject area) worked through a series of "benchmark" portfolios--portfolios that had been selected by assessment developers to represent particular score points. Using these as training portfolios, readers were taught to evaluate portfolios using the same evidence, criteria, and procedures that the assessment developers used. Before being allowed to score new portfolios independently, they were tested on two additional benchmark portfolios. During actual scoring, readers worked first individually to evaluate the portfolios, and then collaboratively in pairs to resolve discrepancies both at the level of evidence and pattern statements and at the level of the overall score.

In the case studies we present, each portfolio was evaluated, completely

independently, by three pairs of readers. The data that inform our case studies include all the written materials they produced, transcripts of their dialogues as they resolved discrepancies, and individual interviews with each of the readers. While we draw from the insights produced by ethnomethodology, our analysis is in no way an example of an ethnomethodological project. Instead of examining how a single pair constructed order over time, which would have been a more ethnomethodological approach, we have taken a thematic approach, exploring the kinds of differences that emerged among them. (Note 7)

To address our questions about how different pairs of readers made sense of portfolios, we condensed and ordered our rich corpus of data through a series of steps. First, a read-through of the data generated a series of themes. These themes were developed inductively--while they relate to the guiding questions that readers were given, we allowed themes to emerge from the notes and dialogue. A file was created for each reader with all the relevant data organized under these themes. Drawing from the data files for each reader, we generated the side-by-side comparison tables that appear below. As we worked, we looked across the data for all three pairs, ensuring that we identified issues mentioned by one pair in the corpus of the other pairs.

In our two case studies we note only a few disagreements between readers in the same pair. In fact, very few clear disagreements emerged within pairs in our data, something that we saw in our earlier work, as well (Moss, Schutz, and Collins, 1998). While certainly there were subtle (and perhaps sometimes important) differences of opinion, these were generally difficult to detect and define in any definitive way. Although readers were told in their training that they were to honor any disagreements between them, in fact our analyses of a number of different efforts to evaluate portfolios using this paired approach has shown that for pragmatic reasons of limited time, among numerous other reasons, readers very quickly accommodate themselves to each others'ways of seeing. (Note 8) Thus, except where there was clear evidence to the contrary, we treated each pair as a unit and integrated comments from both members into a single "opinion." However, we do return to this issue of differences between readers within the same pair at the end of the first case study.

Before we move to the case studies, it is important to acknowledge that in qualitative research it is always a challenge to give enough data necessary to illustrate the particular points central to one's argument without overloading the reader with information. One always draws selectively from a much larger data set. Even in our most extensive case study, #1, we address less than one-half of the issues that emerged in our complete analysis. Given the rich data we have provided about each theme, it should be possible for readers to evaluate our conclusions about how readers generally made sense of the portfolios. We have not provided enough data, however, to fully illuminate why each pair gave the *final* score it did. Our goal is to illuminate the (different) ways readers made sense of the portfolios, even when they attended to the same evidence and seemed to value the same criteria, and to consider the implications of this for assessment practice.

Case Study #1

Our first case study examines readers' evaluations of a portfolio that generated quite a significant difference amongthe pairs in their final scores. Pair 1, Charlene and Iris, scored this portfolio a "one," which was describedas a "conditional" performance. As planned for the operational system, ateacher who received a "one" would not receive a license until she repeated and received a higher score on the assessment. Pair 2, Robert and Sandra scored this portfolio a "three," a "proficient" performance. And Pair 3, Burt and Dani, scored the portfolio a "four," an "advanced" performance. It is important to note that we chose this portfolio for data collection before it was scored because the trainers informed us that it contained characteristics that they felt would make it challenging to score. We return to this issue later on.

We begin with a brief overview of the major components of what we call the Civil Rights Movement (CRM) portfolio and of the general process readers used to evaluate portfolios. We then give an overview of the perspective each reader pair took on the portfolio, discussing key areas of agreement and disagreement. This is followed by side-by-side comparisons of the pairs' views on three different topic areas—"Pedagogy in the Response to Literature (RTL) Section," "Teacher Control and Classroom Dialogue," and "Supporting Different Ways of Learning." (Note 9) These topic areas were chosen because they illuminate most of the key issues that divided the readers. Again, they represent only a sample of a larger number of topic areas that emerged in our analysis.

Portfolio Overview

On the cover sheet, the male teacher who constructed this portfolio noted that his school community was in an urban setting, and that the class he focused on was a seventh grade reading class with 18 students. The portfolio focused on a unit he constructed on the Civil Rights Movement (CRM), and he provided students with a range of non-fiction materials, including pictures, videos, and written texts. The materials presented in this portfolio differed from most portfolios encountered by readers, which tended to focus on literature. Like all of the English/language arts portfolios, the portfolio was divided into two main sections, the first entitled Response to Literature (RTL) which was meant to focus on student engagement with literature and included a relatively small composition component. The second, entitled Writing, was meant to focus directly on a teacher's process of teaching composing. In some portfolios these two groups of lessons were not related to each other, but in this portfolio they were both a part of the same unit on the CRM.

In this case, the content covered in the RTL section included two informational videos and readings on school desegregation. For their assignments, students answered questions about the readings, drew pictures of a scene from the CRM, and wrote a "newspaper article" on something the class had discussed. In the Writing section students wrote a persuasive letter about the civil rights movement to the mayor of their city.

Overview of Each Pair's Perspective on the Portfolio

Pair 1: Charlene and Iris (Final Score of 1). Both readers felt the teacher had

what Iris called “impressive” and what Charlene called “lofty” goals, especially in the RTL section, but that, with the exception of aspects of the Writing section, he could not meet these goals. They both read the early pages of the portfolio with “high expectations” (Charlene) that were not fulfilled in the rest of the portfolio. Although they didn’t think the teacher’s activities fit together well in the RTL section, they thought the Writing section was very coherent, almost a “different teacher.” Iris noted that “if anything, [his] . . . unit had one thing going through it,” the Civil Rights Movement, “as badly as it was delivered,” although it was “not a theme by any means.” Ultimately, the readers agreed that the teacher was basically “task oriented,” and was focused on “skills development and literal comprehension of text.” The pair seemed to feel, as Charlene stated, however, that the portfolio was a “high one:” that he had good ideas, and that he was “headed in the right direction.” Some strong mentoring, she argued, could help him “really bring together the strengths of his work that we see and change some of his overly structured processes.” However, Iris, at least, seemed to question whether the teacher could have actually completed all the activities with his class that his logs said he did. “I couldn’t do it as a veteran teacher with [inaudible] amount of years. I’m not sure this played out. And I don’t want to read between the lines, or investigate. That’s not my job. It’s simply to assess. But . . . it was totally amazing. He seemed to pack a lot [in] during the day [and] I wondered how the information was addressed in a 45 minute period of time.”

Pair 2: Robert and Sandra (Final Score of 3). Both readers liked the teacher’s general goals. Robert argued that although they were not especially “lofty” or “thought provoking,” they went beyond what a “solid basic teacher [score level of 2] would attempt.” They stated that the teacher organized his teaching around the concept of the Civil Rights Movement, although Sandra noted that he didn’t articulate this well in the beginning, so that she thought his focus was Martin Luther King until she got to the Writing section. Both pointed out that he drew from the students’ interests, and noted that the teacher’s pedagogy was “structured and comfortable.” As Sandra said, “in two more years [his classroom] will become clearly a structured and comfortable environment. He’s growing!” Nonetheless, as Robert said, they felt that the candidate did “have moments of that basic two range.” Sandra was impressed by the way the teacher “tied the entire thing together for these kids,” creating an educational experience that was “as good/tight as [one] would expect from a second year portfolio.” She also noted that it was laudable that the teacher had students read newspapers in the classroom, something she didn’t “usually” see.

Pair 3: Burt and Dani (Final Score of 4). This pair seemed to feel, as Dani said, that the teacher “accomplished what he promised to accomplish” and had “clearly stated goals that we felt, I felt, he achieved.” They gave the teacher some credit for teaching a social studies unit and not a literature unit. “I’ve only taught literature,” Dani noted, stating that “when his material wasn’t just faithful to literary text but faithful to historical text, I found that I respected him and had a high regard.” She also “cut him slack” because he was teaching 7th graders. Both Burt and Dani thought the unit was very well tied together. And they felt that the classroom was extremely “comfortable.”

Initial Analysis

Already one can see important differences in how the readers were responding to the portfolio. Pair 1 framed the teacher's goals as "lofty" and complained that he did not achieve them, whereas the other two seemed to think that the goals were simply reasonable (*not* "lofty," in Pair 2's terms), and that he *had* largely achieved them. Pair 1 felt that much of the teacher's pedagogy did not fit together coherently, whereas the other two stated that it was tied well together for students. It is easy to see how each of the pairs' initial response to the teachers' goals might affect their later interpretations. Importantly, the second disagreement about pedagogy seemed to clearly reflect a difference in how readers constructed a "story" about the coherence of the teacher's pedagogy: two pairs were able to make coherent "sense" of what the teacher presented in the RTL section, and one pair was not.

Other differences are also evident. For example, the latter two pairs noted that the classroom was "comfortable," something not noted by Pair 1. Of course, this involves a vision both of "comfortable" and of the kinds of student actions that would indicate "comfort." Also, early in their reading of the portfolio, Pair 1 seemed to raise issues of trust, questioning whether the teacher really did what he said he did, an issue that returns below and something not stressed at all by the other two pairs. Finally, Pair 3 seemed to give the teacher credit both for teaching non-fiction and for teaching seventh graders, something the other pairs did not mention (although Sandra, in Pair 2, did say the use of newspapers was unusual).

In the following sections, we show how differences in the "stories" the readers were constructing about the portfolio emergedas increasingly important sources of disagreement across the pairs. Each section focuses on one of three themes that emerged from the dialogue and interviews as key to the issues dividing readers: Pedagogy in the Response to Literature Section, Teacher Control and Classroom Dialogue, and Supporting Different Ways of Learning. Within each section, we present multiple (numbered) side-by-sides showing the three pairs of readers using the same or similar evidence to develop their own perspectives.

A: Pedagogy in the Response to Literature (RTL) Section

	Pair 1: Charlene and Iris	Pair 2: Robert and Sandra	Pair 3: Burt and Dani
Student Engagement with the Content			
A1	Pair 1 felt that while the teacher said he would get his students to an interpretive level in RTL, the evidence provided little support that he did so. As Iris	Pair 2 acknowledged that the teacher, as Sandra stated, focused on helping students find information and "established a relationship	In Pair 3, Burt stated that the teacher focused on acquisition of knowledge in RTL. Dani felt that the "many prereading and prewriting activities [were used] to empower students with information—background." She noted that the teacher

	said, “he didn’t quite take these kids to the next step.” She saw “very little in the way of student interpretation.” Charlene felt that there were “many missed opportunities” for the teacher to engage students in interpretation. Both readers continually noted “little” deviations, but focused on the larger pattern.	between responding and composing that was literal in nature.” There was, Sandra stated, only “some evidence of elementary analysis and some creative response,” mostly framed around “students interpreting quotes in reference to historical events and what they would do.” She noted that “students read texts and responded to a variety of questions which guided them to finding, and in a few cases interpreting, the facts.” “He gave them case studies about history,” she said, “and research and let them glean the information.”	was “always mindful of the . . . student need to select and grasp and interpret information that’s significant in the text,” and stated repeatedly that the material the teacher was presenting was challenging for his middle school students.

From the evidence above, it is clear that the three pairs agreed on much of the basic “evidence” in the portfolio, even though they often framed these in different ways. They all agreed that the teacher focused on acquisition of knowledge. However, Pair 1 seemed especially concerned about this, noting the teacher’s missed opportunities and the fact that he thought he was teaching critical thinking when he wasn’t. In fact, Pair 1 focused on negatives that they then qualified, while the other two pairs pointed specifically to positive exceptions to the negative pattern while acknowledging, in Pair 2, their limited number, and in Pair 3 the teacher’s focus on finding information (also see row A2, below). Pair 3, however, framed the teacher’s efforts to impart knowledge to students in a very positive light, in that he was helping students pull out the important information, “empowering” them with information, and Dani linked this to the fact that they were engaged with very challenging material. In fact, Pair 2 and 3 both noted that the teacher helped students to, as Sandra said, “glean . . . information,” a more active engagement with the material than that described by Pair 1.

Opportunity for Students' Personal Responses			
A2	Iris noted that there was "little" opportunity for personal response on the part of students. Charlene pointed to the video where the teacher had an opportunity "to take the visual and turn it into feelings," but, "took it to a literal stance." There were, she said, "many missed opportunities . . . [to] elicit student feelings."	Sandra noted that "a few questions asked what students would do." and while students were "directed to pull information out" there were at least a "few questions" that asked them to explore how they "felt." On the video, for example, "up until question three he was being factual in Eisenhower's speech, but four and five ask the students to walk in his shoes."	Pair 3 felt that there were too few opportunities for students to make "personal connections" to the material. They thought this was the key limitation in the teacher's pedagogy in RTL. Dani praised the teacher for asking students how others in history might feel, but "it wasn't bring the civil rights issue into your own neighborhood . . . into the student's own world." Dani was reassured, however, when the teacher asked the students to put themselves in Eisenhower's shoes on the video. From the perspective of his larger performance she stated that "I would presume he would do something with it [personal response] later. Or he was just loosening them up to get them to think and feel what those students felt like, or . . . . But he alluded to it, so I know he's mindful that it's necessary. I just didn't see many instances [in his pedagogy]."

The pattern from row A1 is repeated somewhat in row A2. Pair 1 stated that there was "little" opportunity for personal response, but repeatedly indicated that they did see *some* exceptions, even if these did not seem significant to them. Furthermore, Pair 1 focused on a particular moment when the teacher could have elicited personal responses but failed to do so--one of his "many missed opportunities." Pair 2 agreed with Pair 1 in essence, but did not mention the "missed opportunity" and, again, specifically pointed to the "few" questions that did ask for a personal response. Finally, Pair 3 was also concerned that there were few opportunities for personal responses to the material, and actually went into more detail about the specific kinds of personal response opportunities that they felt were missing than either of the others. In doing so, however, they both stressed the kind of responses that *were* encouraged--imagining how others in history might feel--and pointed to a key example where the teacher did in fact encourage a personal response--the Eisenhower example (which was also noted by Pair 2). Furthermore, Pair 3 noted approvingly that the teacher was "mindful" of the need to encourage personal responses, even though he didn't

do it. Contrast this with Pair 1's criticism in row A1 that even though the teacher said he would get the students to an interpretive level, he didn't do it. For Pair 1, his failure to do what he said is used as evidence against him. This pattern repeats, below.

Student Understanding of the CRM			
A3	Pair 1 stated that the teacher basically led his students to a literal understanding of the material he presented. Both were especially concerned, as Charlene noted, that "he actually thought he was using critical thinking" when he wasn't.	Pair 2 seemed to agree that the teacher led the students, as Robert said, to "develop a critical and analytical response to the material . . . leading students to an understanding of the CRM." The teacher had students integrate the many different materials he gave them through empathy, helping them gain, Charlene said, a "perspective on being black in America then."	Dani stressed that the teacher gave "many . . . opportunities [to] apply students' higher order thinking," and that numerous strategies were used "to develop students as critical/analytical thinkers" producing "independent thinkers who respond and interpret non-fiction with a critical thinking stance and work supported opinion into various activities." She was impressed that instead of the usual "band-aid approach to black history . . . what a way to introduce kids to the reality of a time and scope, the whole history of a country, and *really* bring them to see where we are now."

In row A3, where the pairs discussed the kind of learning taking place in the RTL section, the first evidence emerges in this topic area that the pairs focused on significantly different aspects of the portfolio. But these differences seem tightly connected to their different interpretations of what seemed like very similar data in rows A1 and A2. Pair 1 argued that the teacher's pedagogy led students only to a literal understanding of his material. Again, they were concerned that he "thought" he was teaching critical thinking when he wasn't. In contrast, Pair 2 felt that the students *were* led to a critical understanding of the CRM, but focused on *what* they had learned--the information they gained. They argued that the teacher helped the students bring this complex material together through a process of "empathy." Here, Pair 2 seemed to have a somewhat different meaning for "critical" than Pair 1, and this different meaning seemed to have developed in the context of their response to *this particular portfolio*. The students gained a "critical" understanding of the material by gaining a new perspective--that of being "black in America." Engagement with and gaining an understanding of this complex material seemed, for them, to have involved the development of a particular kind of critical perspective. As usual, Pair 3's stance was the farthest from Pair 1's with Dani's conviction that

the students engaged in “higher order thinking” and became “independent thinkers.” It is hard to tell how Pair 3 came to this conclusion given their statements in the first two rows. However, their reasoning may have been similar to Pair 2’s, given their earlier focus on the difficulty level of this material for seventh graders and the way the teacher “empowered” students with information. Again, the disagreement in this row may relate to the pairs’ earlier differences over whether the pedagogy in the RTL section was coherent or not: Pair 1 seeing it as incoherent, and Pairs 2 and 3 seeing it as tied together well and thus able to help students achieve a coherent understanding.

In summary, while there are quite significant differences between how each of the pairs evaluated the RTL section, most of the differences between them appeared to result from the ways they constructed “stories” about what was happening in the teacher’s classroom: focusing on different aspects of the portfolio and framing teacher actions in different ways. Further, it is already possible to see in this initial example the ways in which a particular framing of the classroom with respect to one aspect can affect the way a pair frames other aspects of the portfolio they encounter.

B: Teacher Control and Classroom Dialogue

	Pair 1	Pair 2	Pair 3
Teacher Control			
B1	Iris and Charlene argued that the discussion video was crucial evidence showing the teacher over controlled the classroom. For Charlene, for example, seeing “question and answer” in the video, “colored all the daily logs where he said ‘discussion’. . . . I began to wonder whether or not those were discussions.” She noted that the teacher asked questions and then answered them himself and that he took students’ literal answers and elaborated them “into	Robert and Sandra felt that the teacher was too directive in the classroom. Sandra noted that the discussion on the videotape was primarily “Q and A,” and stated that the teacher did not “fully understand, as evidenced on the video, that discussion is [not supposed to be] teacher-directed.”	Dani noted that the “discussion” was really just a “question and answer session” something that really “annoyed” her: “It’s like, does anybody know how to have a real discussion?” “How,” she wondered, “is this empowering them [his students] as . . . members of a discussion?” But while she noted that the teacher didn’t have “discussions,” “he at least engaged most of the class, and they’re ‘what would you do?’ questions, they’re ‘what about. . .?’ questions. They’re not just literal comprehension. So I had to give that though too.”

	perhaps a more interpretive level" himself. Charlene was especially concerned that the teacher didn't understand that question and answer didn't count as a discussion.		Pair 3 specifically noted the exception of questions that asked students to "walk in Eisenhower's shoes."

In row B1, the pairs seemed generally to agree that the teacher over controlled the classroom and that he didn't understand what "discussion" meant, given the conceptual model of this assessment. Again, however, Pair 3 went farther than Pairs 1 or 2, noting specific strengths of the teachers' dialogue that the other's didn't. For Pair 1, which already had questions about the accuracy of the teacher's logs, the teacher's apparently different understanding of discussion raised even more questions (a good example of what we referred to in an earlier paper [Moss, Schutz, and Collins, 1998] as a "cascade"). While the lack of "dialogue" in the video must have also colored how the other pairs read the term "discussion" in the teacher's logs, they did not explicitly raise this as an issue. And, again, Pair 1 was more specific about what the teacher was doing in his "discussion" that didn't fit with their conception of a discussion, while Pairs 2 and 3 had brought up the positive Eisenhower example that Pair 1 didn't mention.

Student Interaction in Peer Critiques			
B2	Pair 1 felt that the video where students critiqued each other's papers in the Writing section "showed tremendous insight." But both were concerned that even here, he didn't seem to be able to let go. In fact, Iris questioned "whether he had to go to that level of facilitation in March if this had been a pattern	Pair 2 thought that the peer critique video showed, as Sandra said, that the students were "comfortable making suggestions." They were impressed with how the students critiqued each other, noting that their language was more sophisticated than most seventh graders: "They said 'you could make this better by doing this,' not 'this stunk.'" Sandra acknowledged that some people might not agree with her, but stated that she liked that the teacher had read the students' papers ahead of time and	Dani was especially impressed by the teacher's ability to "stay out of their territory" in the peer editing video. "When a teacher can take part in that editing process and not take the pen . . ." she noted. She felt the editing process "would empower [students]. And he had the proof" in the peer editing video. "They were being honest about their [opinions] . . . for such young students they were being very evaluative."

	ingrained in those kids from a very early timeline. So I don't want to get suspicious about motives. I don't want to go there."	helped to facilitate the peer conference. "He knew specific questions he wanted to ask about each students' paper to draw the peer editor towards making conclusions about the work" which is "something you rarely see in teachers." They noted, however, as Robert said, that "he lets the kids peer edit for themselves and then he won't let [go]. . . . It's a mixed message on that."	She noted that "he exhibited a lot of trust in the writing process with them. I have no doubt that's what made me buy in because I saw so much trust there in their process of writing." It was clear to Pair 3, as Dani said, that "this isn't the first time" his students had engaged in peer critique. "I thought that was quite a height [the teacher] had climbed . . . and [he] deserved that regard."

In Row B2, the readers agreed that the peer critique video was extremely impressive, although Pair 1 was much less specific in their praise. However, Pair 1 emphasized that the teacher'sfacilitation, while effective, showed the his inability to give up control. Furthermore, even though Pair 1 seemed to agree with the others that the actual activity of the students seemed to show that they had learned how to engage in peer critique, the teacher's facilitation appeared to raise questions for Iris about whether the teacher really *expected* the students to be able to do it. Thus, this raised questions for her about whether the teacher really had students engage in peer critique regularly in his classroom. Although she didn't want to get "suspicious," she nonetheless *saw* potential issues with his "motives" in facilitating. Pair 2 also noted that the teacher's facilitation was an indication of his need for control. But while Sandra acknowledged that it would be possible to downgrade the teacher for actively facilitating what would normally be a more independent peer activity, she felt that in this specific case he facilitated very effectively in ways that showcased some of his skills as a teacher. Finally, Pair 3 framed the teacher's activity in an entirely positive light--extending on Sandra's positive statements by focusing on the teacher's ability to *resist* taking control *even though* he was facilitating. Thus, in contrast to the others, for Pair 3 what appears to be essentially the same data from the peer critique video gave direct evidence of the opposite of what the other two pairs saw, of the teacher's ability to trust his students and *give up* control even when he was in a position where one might expect him not to.

Variety and Quality of Activities			
B3	Iris and Charlene were	Pair 2 noted, as Robert said, that	Although Pair 3 was impressed by the variety of activities the

	impressed by the variety of activities the teacher provided, but felt that his students had "come to over-rely on him," and noted that he was unable to "break away as a center of control" even though "he himself admitted to giving too much of himself."	while the teacher "provided a variety of opportunities" for learning, "they were very teacher directed." And they agreed that, as Robert noted, he "provided a positive and challenging learning environment, except that he stepped in and gave them the answer. . . . He realized they have frustrations, so he steps in and says, let me do this for you."	teacher provided, Dani noted that "it is his choice in how they would show off that material." "Even though he was controlling in the process," Dani noted, however, "I did feel that students were empowered through the process." She was especially impressed by the "million strategies" he gave them, noting that he made sure the students had "all of the things that would help [them] build to the goals required in class." Burt agreed that the "teacher . . . frequently directed." Repeatedly, Dani focused on the distinction between the teacher's control over the structure of the assignment, and his working towards their "independence" "in the reading . . . the process of writing, [. . .] the process of getting to an end result in a piece of my writing. I have the power to be in control of it. And that's what I think he gave those students." They agreed, however, that his writing process "was not a flexible plan," noting that it was "articulated, but still controlled." Finally, both felt that the teacher would use less control if he taught a novel instead of non-fiction material.

Row B3 extendson the pairs' concerns about teacher over-control. Again, the pairs seemed to use very similar evidence to very different ends. For Pair 1, the teacher's statements about his need to pull back were "admissions" of his inability to do so. While Pair 2 acknowledged the teacher's directiveness, they *saw* his "frustration," the well-meaning intentions behind his difficulty in giving up control. And, while Pair 3 also saw the teacher's directiveness, Dani stressed aspects of the teacher's pedagogy where the students did have some control. Finally, Pair 3 argued that the non-fiction material used in this particular portfolio may have required more control than fictional material, and that the "strategies" he gave them helped them "build to the goals required in class"--thus aspects of the teacher's control in *this case* may have been appropriate.

In this topic area, then, the pairs tended to cite much the same data (although there were differences in focus). Further, they appeared to agree on the definition of an effective discussion, and on the importance of teachers' giving

up control of aspects of their classroom. Most of the disagreements, then, appeared to result from the different "stories" the readers were constructing about what was happening in the classroom.

C: Supporting Different Ways of Learning

	Pair 1	Pair 2	Pair 3
Teacher's Expectations of Students			
C1	These readers agreed that the teacher had low expectations for his students. As Charlene said, he "really underestimated" them, not because his goals were low, but "in the way he taught. . . . They meet his expectations because he doesn't do anything to raise them." As Iris noted, "he expected the low level learner was going to be a little overwhelmed, no matter what the opportunity." And as Charlene stated, "my gut response was all of these judgments about 'they can't get it,' or 'I paired them with somebody else so that they can just sit back.' That just kind of concerned me."	This pair felt the teacher was very responsive to the different ways his students learned. Sandra "found his instructional strategies as a reading teacher . . . [to be] designed to support the [students'] development as readers . . . . He mentions . . . the slow learners, low level readers, which indicates that he has some IEPs that he's working with in the class. . . . I thought he did a really good job." They especially noted his effort to pair stronger and weaker students.	Dani noted that "his mind's eye is always on his slower reader, the less achieving student, always working the crowd to empower those readers into being equal." "He always went . . . toward those students who might have slipped through the cracks. And he was constantly committed to ensuring that they . . . didn't." Burt noted that "he says a lot here about slower kids and he is very cognizant of their needs and trying to find ways that he can help them. But yet he's also looking at the brighter kids." Both saw, as Dani said, that he was "mindful of the whole class . . . as individuals."

In Row C1, Pair 1 seemed to interpret the teacher's repeated mentions of low level students as indications that he was "underestimating" them. In contrast, Pairs 2 and 3 felt that many of the same statements showed his real concern for the very same students. As usual, Pair 3's statements appeared more glowing than Pair 2's.

Teacher Knowledge About Students

C2	Despite their discomfort with the teacher's low expectations, Pair 1 acknowledged, as Iris noted, that "he really did hit on the kids' interest levels [and] their innate ability, the effort produced, and the thing that I bought into was the fact that he said he was helping these children in the classroom and then having them stay after school and work with him." As Charlene noted, "he's aware of student needs." And Iris stated that "teacher's knowledge of student interest levels, ability, and effort produces some teacher recognition of [the need to] adjust instruction [and] support student needs." Iris acknowledged that the teacher "drew useful conclusions and made decisions regarding future practice." Both readers noted the teacher's individual work with the ESL student, which Charlene said showed potential and that "some serious mentoring" could "really bring together the strengths of his work that we see and change some of his overly structured practices." At the same time, however, Charlene wondered "whether he [conferences with students] regularly," later stating, "I won't even go into the fact that I think she [the ESL student] came after school, not that he suggested it." Nonetheless, she noted "in his defense [that he made a] lot of comments regarding the students. . . and he mentions pairing [the ESL student] up with a friend of hers who works with her after school. [So] I'm not saying he has no expectations, or that he is damaging students, but it	The readers agreed that, as Sandra said, the teacher "used his knowledge of student needs to inform the types of activities he selects, his monitoring and pairing /grouping of students" actually modifying instruction "so that all students are successful on some level." Robert noted, "he does all kinds of things . . . that invite other than just what we consider perhaps that traditional type of learner into the picture." They specifically noted his work with the ESL student.	Both readers cited a wide range of evidence indicating that the teacher was paying attention to the differing needs and ability levels of his students. They were impressed with how he dealt with his ESL student, and thought he was, as Burt said, "very knowledgeable about the students." Dani noted that "he was always mindful of making sure that the student who most likely would not build to the right end result had all of the things that would help build." In his planning, she said, he was "always mindful of the groups' impact on their time and individual students to understand and interpret."

	seems like there is a pervasive low expectation of students." There was, the two agreed, "some opportunity to learn for some."	

In Row C2, all three pairs acknowledged theteacher's knowledge of his students and the efforts he made to help them. However, in row C1, Pair 1 had already classified some of the teacher's pairing efforts as evidence of his low expectations. So Pair 1 had fewer examples of positive teacher efforts to cite in row C2 than the other pairs. And, again, Pair 1's lack of trust led Charlene to raise questions about whether the teacher generally helped all of his students given the way the teacher phrased his description of his efforts to help his ESL student, downgrading the importance of what was a key example for the other two pairs. Nonetheless, Pair 1 did use much of the same evidence that the other two pairs used to show that the teacher did an effective job helping his students, and this mitigated and nuanced their statements in row C1 about his low expectations. Thus, Pair 1 noted that there was "some opportunity for some" in this classroom.

Again, most of the disagreements in this topic area appeared to arise not out of differences in criteria or in what they saw, butout of differences in the "stories" the pairs constructed.

Differences on the Final Score Among Readers Within the Pairs

Up to this point in this paper we have avoided discussing differences *within* the pairs over interpretations of the portfolio. However, we know from the readers' comments in individual interviews separate from their partners that there were some differences in what they said they focused on in reaching their final interpretations. Given limited space, we will not go into these in detail. However, what is interesting is that four of the six readers acknowledged in their individual interviews that the CRM portfolio might deserve a different score than the one they had given it. Importantly, these differences seemed to reflect not disagreements with partners but instead individual reader's uncertainties about the correct score for the portfolio. In Figure 2, we plot out the range of different scores members of the pairs said they might be willing to give the portfolio.

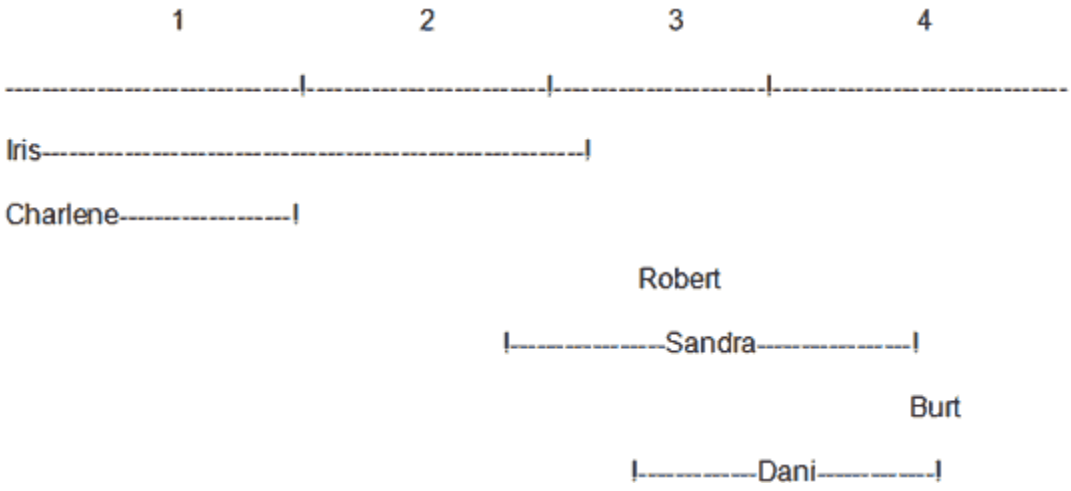

Figure 2: Differences within pairs on Civil Rights Portfolio (Case Study #1) (Note 10)

In Pair 1, Iris argued that she didn't think "enough evidence was given." If she had been given more evidence that had shown that the teachers "goals had been met" then she might have given him a two, "maybe even a three, depending on the level of the evidence." Her partner, Charlene noted that the teacher "had good ideas and was headed in the right direction," and acknowledged that she felt the portfolio was a high "one," or almost into the "two" category. In this pair, Charlene was clearly the junior member, with much less teaching experience than Iris, and it is possible that in a more equally matched situation she might have been willing to go higher. In Pair 2, Robert seemed to stand fairly solidly on a score of "three" for the portfolio, even though his statements in dialogue seemed to indicate more flexibility than this. Sandra explicitly noted that she struggled between a three and a four, and at one point seemed even to be playing with the possibility of giving the portfolio a two. In Pair 3, Burt stated that he felt very "confident" about his score of 4, but Dani had argued in their dialogue over whether it was a three or a four, finally becoming convinced by Burt's evidence that it was a four (although her interview indicated she still may have had some questions).

Thus, most of the evaluators of this portfolio were not very confident that they had arrived at the correct final score. We return to this issue in our discussion of the second case study portfolio, where we found very few differences between the readers.

Overall Analysis of the CRM Case Study

If our presentation and discussion of the data, above, has been at all convincing, it should be clear by now that most of the differences amongthe pairs involved "stories." Again, it is important to emphasize that the aim of this case study was not to determine the direct cause of the pairs' final score differences. Instead, we seek only to show that "story" issuescan "reasonably" be said to have reflected a pervasive source of differences between the two pairs. And while we cannot know whether or not disagreements over stories caused the final discrepancyamongthe pairs on the score they gave a portfolio, our goal is only to show that it is an important enough source of problems that

one could *imagine* it frequently resulting in such a difference in final scores. With respect to this particular question, then, what we have tried to do is crack open what is often treated as the "black box" of reader judgment processes.

We argue that the three pairs of readers in this case study seemed relatively well trained in a conventional measurement sense, in that they generally seemed to share a perspective on what counted as effective teaching and they were generally able to cite much the same "evidence" from an extremely complex portfolio under relatively severe time constraints. In fact, it is hard to imagine readers doing much better. Given the complexity of portfolio assessment, it seems impossible that readers will always cite exactly the same data--even if they are generally constructing the same "stories." While readers could always be trained better, the skills these readers exhibited seemed at the very least what we can expect from adequate readers.

Nonetheless, these pairs disagreed quite substantially about the performancelevel of this portfolio. Their final paired scores covered the entire possible range of scores--from 1 to 4. If we are right that these readers seemed at least adequately prepared to evaluate these portfolios, then it is reasonable to assume that the problem of differences over "stories" may be significant for other evaluators who face similar challenges when encountering similar kinds of portfolios.

Importantly, the evidence does not seem to indicate that any of the readers made an inappropriate rush to judgment early in their evaluations of this portfolio. In fact, as the data we present indicates, all three pairs were careful to acknowledge conflicting aspects of this teacher's performance, finding a range of aspects in the same areas. Instead, it seems clear that their continuing observations were influenced in a range of ways by observations they had already made as they moved between their vision of the whole and their understanding of the meaning of a particular piece of data. Further, because the readers had limited time to complete their evaluation, they had limited attention to focus as they went through the portfolio. Instead of a rush to judgment, differences in focus appeared to reflect evidence-based decisions about where readers would focus their limited energy, representing perhaps unavoidable choices whenever evaluators must choose what to foreground and background in their reading of a collection of material.

In our next case study, we present an example of a portfolio that did not seem to elicit significant differences over the correct "story" for the portfolio. First, however, we address a concern raised by others who read initial drafts of this paper: reader inferences about teacher intentions and motivations.

Discerning Motivation and Intention

Some of the "story" disagreements visible between the pairs involved interpretations about the motivations or intentions of the teacher. But is this a reasonable area for readers to investigate? Shouldn't they be evaluating only what they can "see" directly and not what they have to infer? Conversely, is it possible (given what we have said about the assumptions of ethnomethodology, above) for readers *not* to raise these issues?

First of all, we argue that these inferences are not fundamentally different from the other kinds of judgments readers are asked to make. Recall Garfinkel's (1967,2002) argument that all language is "indexical," pointing to some state of affairs that it cannot entirely capture. And we showed these kinds of inferences about the teacher's classroom being made continually by the readers in the CRM example. Readers had to decide whether the teacher's pedagogy was "tied together," how important exceptions to a pattern were, and what should "count" as "critical thinking" in the context of a particular unit. In these and many other cases, the readers had to infer, from the limited data they had, the general state of the larger classroom. Other studies of reader processes have shown much the same thing.

Even if we could eliminate questions of motivation and intent, however, it is important to understand that this would make it impossible for readers to ask *why* a teacher made a particular move or statement. Yet current teacher standards make processes of teacher decision-making a central part of good teaching. Thus, readers often struggled to understand why a teacher was acting in a particular way. A good example came in an interaction between Robert and Sandra in a discussion of why the teacher in the CRM portfolio had failed to grade some assessments. Robert said,

I don't know whether he's consciously saying 'they're going to think this is bad because I don't assess [these papers], but I'm going to do it anyway,' or, 'I said this is what I'm going to assess, this is what I'm going to assess, and it's got to stand on it's own' . . . . I don't know, but it's either gutsy or stupid.

To which Sandra responded, "Maybe he's worn out by this . . . class." In other words, Robert was trying to decide whether the teacher had made a conscious, calculateddecision based on his pedagogical approach,or whether he just didn't care. And Sandra offered a third option—that this specific case was an exception for this teacher because at this point he was just "worn out." The pair ultimately agreed that the teacher had made a conscious decision not to grade the assessments based on his pedagogical strategy, and thus gave the teacher credit for his process *even though they did not agree with his final decision*. In more subtle ways, issues like this frequently arose amidst readers' efforts to evaluate. It is difficult for us to see how readers could avoid making such judgments and still operate under the standards for effective teaching that currently predominate in the field.

On a more basic level, assumptions about intentionality are required when we judge whether actors are responsible for particular actions or events. For example, remember when Charlene, in the CRM case study, questioned whether the teacher made a practice of working with students after school. In that case, she pointed to evidence that the ESL student approached the teacher seeking extra help, and not vice-versa—in other words, she emphasized that the *student* and not the teacher was the agent in this case. Deciding between these options involved understanding the teacher's motivations for his actions. More broadly, what would happen to the validity of the assessment if we did not ask readers to differentiate, for example, between an accident and an intentional act?

In fact, evidence from ethnomethodological and other studies (e.g., Gibbs, 1993; Schegloff, 1999) indicates that it would be difficult to prevent readers from making inferences about teacher veracity, intentions, and the like, even if we wished to. As we noted above, even when Garfinkel attempted to create situations where there was no motivation for his actions internal to the context he was acting in (replacing one white pawn with another one) people had great difficulty in believing that no motivation was involved. In Garfinkel's work "deviations from . . . [normal] sense-making procedures were instantly interpreted as 'motivated' departures on the part of experimenters who were treated as acting from 'special' if presently undisclosed motives" (Heritage, 1984, p. 99).

Furthermore, ethnomethodology argues that the very meaning of any statement always involves "grasping the *purpose or the motive* for its being produced" (p. 101). Even the seemingly simple act of describing a situation raises questions about intent. As Schegloff (1999) showed, there is no such thing as mere description since there are a potentially infinite number of ways any particular situation can be described. A description always involves some answer to the question "why that now?" Thus, Heritage (1984) arguedthat any description of a complex setting or interaction is necessarily "selective in relation to the state of affairs it depicts, . . . [so that] part of the process of" understanding an action or a statement necessarily involves "grasping the purpose or motive for its being produced at a particular moment" (p. 151). Other experiments in psychology also show the centrality of interpretations of intention to human interaction. For example, studies indicate that our memory processes are highly influenced by what we understand as an intention behind a statement. As Gibbs (1993) noted, it is often the case that "the [presumed] intention behind the speaker's utterance is encoded and represented in memory, not the sentence or utterance meaning" (p. 187). And "experiments show that infants as young as nine months begin reliably to interpret certain behaviors as intentional (e.g., pointing)" (Richards, 2002, p. 2). When we cannot grasp someone's purposes or motives, Garfinkel'sand other's work indicates, we often find ourselves disoriented. (Note 11) Ultimately, studies like these show that in our moment-to-moment interactions, when people act we treat them as agents and "see" motivations.

And we often saw readers "seeing" such motivations as an apparently automatic part of reading the portfolios. For example, in one case Sandra first thought the CRM teacher's unit was about Martin Luther King, but then *saw* that it was about the Civil Rights Movement. In another case, she first thought the teacher had used multiple activities to support slower learners, but then "realized" that he was using multiple activities to engage the multiple intelligences of all of his students. Another fascinating example arose when Iris, discussing the peer critique video in the CRM case study, questioned "whether he [the teacher] had to go to that level of facilitation in March if this had been a pattern ingrained in those kids from a very early timeline. So I don't want to get suspicious about motives. I don't want to go there." She didn't want to get suspicious, but she *was*, she didn't want to discuss motives, but she *saw* them nonetheless. As with the Duck/Rabbit example, such active sense-making is our default approach.

We now move to a discussion of our second (abbreviated) case study.

Case Study #2

For our second case study, we provide an example of a portfolio that contrasts with the CRM example, representing far fewer disagreements among the pairs. Unlike the CRM portfolio, which the trainers had identified as a difficult-to-score portfolio, this portfolio was chosen by the trainers as a "benchmark" portfolio, as a clear example of a score level. And, in fact, the three pairs of readers agreed on a score of "one," constructing strikingly similar stories about the portfolio. (Note 12) The possibility that readers can be prepared to distinguish between more and less ambiguous portfolios is an issue to which we return in our conclusion.

Because there were few examples of "story" disagreements in readers' evaluations of our second case study portfolio, we provide only a short, schematic discussion of it, here. It is important to note, however, that the same careful process of data analysis used in the CRM case study was used here as well.

The teacher in this second case study portfolio organized his instruction around a famous young adult novel, engaging his students in a range of activities, including a mock trial of one of the characters. The similarities in the stories constructed across reader pairs included the following: The pairs felt that the teacher in this portfolio controlled the classroom to the point that the children had little or no choice in their activities. They worried that the teacher seemed concerned about getting his students to do exactly what he told them to with little focus on whether the students were actually learning anything. And they felt that dialogue in the classroom largely fit with the teacher's pattern of control, involving simple question and answer, with little real "discussion." In general, then, the readers agreed that while the teacher'sactivities had potential,he had no clear purposes for them. The readers also thought the teacher had low expectations of his students.

As with the CRM portfolio, we did see places in the data where divergence between the pairs *might* have been possible. For example, at one point Sandra seemed to treat the teacher's mistakes in his own writing in the portfolio as an "editing" problem, whereas the other readers interpreted his mistakes from the beginning as an indicator of problems with his knowledge of English grammar. This raised the possibility that Sandra might infer that the teacher did not have problems, himself, with issues of grammar. However, she saw similar grammar mistakes on the videos when he was working with students. Thus, along with the others, she eventually read the teacher's mistakes in the portfolio as part of a pattern of evidence that the teacher lacked some basic knowledge of English. In this and in other examples, in contrast with the CRM example, the preponderance of evidence seemed to prevent initial ambiguities from leading to significant disagreements.

On only a few generally subtle issues did disagreements between the pairs seem to emerge. In fact, the only major disagreement about the teacher's

performance involved interpretations of the teacher's ability to control his classroom. Pair 3 saw "strengths in classroom control," while the other two pairs specifically noted his lack of ability in this area. In addition, while Pair 2 simply noted his inability to control his class, Pair 1 went further and pointed out that he "didn't alter his approach when" faced with a "chaotic" environment and when "the kids were not doing things." For Pair 1, then, the teacher's response to the "chaos" in his classroom also indicated the teacher's inability to change his instruction in response to challenges, something not noted by the other pairs.

Other more subtle divergences between the pairs included their interpretations of the "mock trial" and of the teacher's understanding of his students. First, with respect to the mock trial, Pair 1 indicated that the teacher didn't know how a real trial works (and should have done some research to find out). Pair 3, in contrast, didn't see the accuracy of the trial procedure as an issue, and in fact indicated that the teacher should have added some "bells and whistles" from *fictional* examples of trial procedures (like Perry Mason). Finally, Pair 2 did not mention the accuracy or the richness of the trial activity specifically at all, instead noting more generally that the teacher's activities focused on helping students learn details that did not relate to the "real world." Second, regarding his understanding of his students, Pair 1 acknowledged that the teacher had some awareness of where students needed to improve. Pair 2 seemed to go beyond this, however, arguing that he "does know his students and their abilities." And Pair 3 went the farthest, noting that he made "astute observations of student behavior."

None of these specific differences in interpretation, however, seemed to trigger particularly significant disagreements in the general "stories" that readers constructed to make sense of this portfolio. The way we interpret this is that the overall "story lines" for this portfolio were robust enough to absorb a few specific differences in inference about the teacher and his classroom.

In fact, we noted a number of examples in the CRM case study, above, where readers from Pairs 1 and 2 seemed to struggle to decide what the correct story line was. Thus, in the CRM case study, many readers indicated at points that the evidence was contradictory enough to lend itself to a range of different patterns. Examples like these tended not to occur in the YAN case study.

Furthermore, we noted that many readers did not seem very confident about their final score for the CRM portfolio. And, in fact, the CRM portfolio was explicitly chosen for us by the trainers as difficult to score. On the YAN portfolio, however,readers generally fell solidly in the "one" category, (Note 13) and the trainers chosethe portfolio as a "benchmark" to be used for reader certification purposes, a portfolio especially chosen to lie clearly within a particular score range. Thus, both the readers and the trainers seemed able to tell the difference, at least in these cases,between portfolios that can and cannot be straightforwardly scored. The YAN portfolio appeared to represent one of many portfolios which tend to produce a single "reasonable" story, and thus a single "reasonable" score, in independent reads. In contrast, the CRM portfolio represented a portfolio that appeared to lack a single "reasonable" story or set of stories. While more reads would help establish more definitively the status of each portfolio, our ultimate goal, here, is not to arrive at confident answers for

these particular portfolios. Instead we seek to show thatour distinction between “reasonably” scoreable and unscoreable portfolios is one that makes sense more generally for the assessment field.

Implications

In this paper, we have discussed two case studies that illustrate how portfolio readers engage with evidence in their efforts to understand a teacher’sperformance. Our findings are consistent with the small body of other literature on readers’ processes in large-scale assessments. We have grounded our analysis in a large body of research around narrative theory, discourse analysis, ethnomethodology, and other fields that indicate such efforts to construct coherent "sense" are an unavoidable and ongoing part of our everyday activity. While we have not presented a large number of cases, we argue that our examples, in conjunction with relevant work in assessment and on processes of human understanding,make it difficult to imagine how readers could *avoid* constructing “stories” in their efforts to evaluate portfolios. In what follows we discuss possible implications that this challenge of “stories” might raise for assessment.

As the CRM Portfolio example indicates, even when readers generally agree on evidence and on relevant criteria,they can construct different "stories" about the teacher's practice. Conventional assessment practices, focusing on interreader reliability, seem unlikely to illuminate these problems. For example, two out of three reader pairsagreed that the CRM portfolio reflected a relatively strong performance. In our in-depth reading of field test portfolios we have encountered other portfolios that resemble the CRM portfolio in their apparently conflictual evidence, yet these cases only sometimes produce discrepant final scores.

Further, as we noted at the outset, the usual approach to solving such discrepancies, focusing on techniques likely to improve interreader reliability on final scores, risks further masking what we argue represents the underlying challenge of ambiguous cases (e.g., Kane et al., 1999;Swanson et al., 1995; Klein et al., 1995;Nystrand et al., 1993). These techniques include:

- further standardizing the tasks, thus constraining the responses that readers encounter;
- disaggregating the portfolio into separate tasks that can be scored one at a time;
- having different readers score the different tasks;
- developing more explicit criteria, including stipulating or reaching a priori agreement on how particular issues should affect scores (e.g., weak commentary but strong performance); and
- separating criteria into separate scales that can be separately scored.

Because all of these practices work to fix or strip context from the information available to readers, they insulate themfrom uneven, conflicting, or otherwise ambiguous responses. While all of these practices can and have improved interreader reliability, they don’t make the problem of ambiguous evidence go away. They simply relegate it to a priori decisions, standardized responses, or

statistical machinery which combines scores according to predetermined algorithms.

It is important to acknowledge that high quality assessment programs do generally have statistical routines that attempt to detect unusual patterns of scores (for raters, for tasks, and for students) and to flag them for additional attention (e.g., Engelhard and Myford, 1994; Engelhard 1994;2001; Wilson and Case, 1997; Myford and Mislevy, 1995). These are important and useful tools. But the routines are only as effective as the information on which they are based. More needs to be done to flag ambiguous cases. In fact, we are increasingly convinced that it is crucial for large scale assessment programs to build in information-gathering processes that are more likely to illuminate the kinds of ambiguities we have illustrated.

The need for such processes is made only more imperative by the fact that most studies that compare alternate scoring practices use consistency between readers'scores and efficiency as criteria for deciding among these practices. Few studies look at broader questions of validity, including the relationship between the assessment in question and other indicators of similar performance. Thus we should be alarmed, but not surprised, by the findings of a study reported by Swanson, Norman and Linn(1995) from the health related professions. Swanson and colleagues reportedthat that the scoring method which producedhigher reliability actually led to lower correlations with the criterion of interest, indicating lower validity. The paucity of studies like these results, in part, from the fact that routine and legally defensible practice does not require assessors to conduct them, even though many measurement professionals wish it were otherwise (NRC, 2001; Madaus, 1990;Haertel, 1991). Without access to such comparisons, it is impossible to know for many large scale assessment programs how choices made by developers affect the meaningfulness of interpretations produced by their system. In other words, we don't know how these decisions affect validity.

In what follows, we argue that it is possible to develop procedures for routinely illuminating ambiguous portfolios and that there are ways that this information could usefully serve efforts to achieve ethical, valid assessments in high stakes settings. It is important to note that most readers, as well as trainers, appeared to know that the CRM portfolio was difficult to score. In fact, it was the trainers' initial identification of it as challenging that led us to collect data on it in the first place. Much the same could be said of the YAN portfolio. It was specifically chosen because the trainers, who selectedit as a benchmark portfolio, informed us that it was relatively straightforward, something also indicated by the response of the readers. In fact, there are a range of indications that evaluators and assessment developers can distinguish, at least in some cases, between portfolios that are more or less easy to evaluate. Indeed, the National Board routinely selects both benchmarks--clearly illustrating a score point--and "training portfolios"--intended to illuminate particular kinds of problems (ETS, 1998).

Controversially for us, however, the National Board assessor training script indicates that all of these portfolios, including the problematic "training" ones, are given predetermined scores to use in the training. (Note 14) The goal of

using training portfolios, it appears, is to teach readers how to score portfolios like these in a consistent fashion. Of course, this makes perfect sense if one assumes that (almost) all complete portfolios have or embody a particular score that can be determined. But if we are right and some portfolios contain ambiguities and contradictions that make it difficult or impossible to "reasonably" assign them a single score, efforts to always train readers to "correct" scores may actually be counter-productive.

Instead, we think it is important to at least explore the possibility that distinguishingin some cases between portfolios that are easier or more difficult to "reasonably" score may provide an opportunity for changes in the ways readers approach portfolio evaluation. In fact, the pressure our readers felt to find the "correct" score for the CRM portfolio may have generated some of the struggles theyfaced. We wonder what would happen if, instead of training scorers to find a “correct” score, readers were also prepared (as trainers are) to identify ambiguous portfolios. Readers might engage in such an effort, for example, by purposefully seeking disjunctions between pieces of evidence or by trying out alternative interpretations of the existing evidence. Such a search for counter-interpretations is fully consistent with good practice in validity research (e.g., AERA et al., 1999; Messick, 1989) and in qualitative research as well (e.g., Erickson, 1986). In our terms this would be akin to seeking out alternate yet equally convincing “stories” for a particular portfolio. Documentation of unevenness or other sorts of ambiguity within a particular portfolio might be institutionalized by providing opportunities for indicating this on the scoring document. Evidence and general observations placed in this new category wouldn’t necessarily preclude giving a final score that represents readers’ best interpretation of the preponderance of evidence. But it would allow them to indicate when issues of ambiguity made it difficult for them to arrive at a final score, and it would allow the portfolio to be flagged for further study--if not during the operational scoring process, perhaps at least as part of ongoing research intended to improve the system.

What, then, might be done with portfolios that are flagged by readers or researchers as ambiguous in an operational system? We can imagine a number of options including the (not unreasonable) option of doing nothing. First, such portfolios might be put into a process that involves a deeper, more comprehensive, review than conventional scoring allows. In this supplementary process, readers might engage in a deeper reading to see if they can surface a coherent interpretation or story that could support a clear final score.

Another option might allow the readers to request more data from a candidate. Perhaps, for example, candidates with particularly difficult portfolios might be asked to interview directly with the readers. Or additional documentation of classroom practice might be required. Of course, this raises all kinds of difficult challenges with respect to time, resources, and simple feasibility. Further, Garfinkel’s work, for example,indicates that it is not at all clear that more data would actually help—as the example from Heritage (described above) illustrated, it could just complicate the issue more. (Note 15) At some point, we simply need to agree that we have “enough” data since no amount of data is ever going to adequately answer all the questions we have. Furthermore, unless we are careful,additional data may actually be misleading or problematic, since

its collection may not be framed as carefully as it is in the larger established system. Despite these challenges, however, in some cases more data might allow readers to achieve more valid evaluations (see Moss et al., in press, for examples of how case studies might complement portfolio-based evidence).

A third approach to this problem would involve shifting portfolios identified as difficult to reasonably score to a larger committee that would take on the responsibility for making an ethical decision given the available evidence, the standards, and the potential consequences of a decision for the candidate and the educational system (similar to what we have argued in Moss and Schutz, 2001). Instead of asking committee members to see exactly as the trainers have asked them to see, they might take on the role of representatives of the larger professional community. Members could take into account the risks and possibilities entailed in the particular performance they are evaluating and make a judgment as proxies for this larger community. Indeed, in another ongoing study with beginning teacher portfolios,we've presented portfolios to expert readers who have not been constrained (as evaluators within an ongoing assessment system are) in the kinds of interpretations they can draw. Under these broadly open conditions, we have asked them to make a pass (achieves regular license) or fail (repeats assessment) decision on each portfolio based on the evidence they see. We have then brought them together to discuss their decisions. In all three cases where we have attempted this, these unconstrained discussions of the risks and benefits of a particular decision, or of trying to reach a decision at all, have expanded to include the needs of the larger system as well as the candidate (Moss, Schutz, Haniford, Coggshall, and Miller, in preparation).

A final alternative we can imagine is that ambiguities noted within portfolios might be ignored in the operational system and used only to inform an ongoing research agenda designed to improve the operational system. There is no question that efforts to note ambiguities will flag more performances as problematic than are currently identified by discrepant or otherwise atypical scores alone. Clearly if they must all be resolved through a more time consuming process, costs will increase, perhaps more than the system can afford. No assessment system (whether large scale or based in an intense, deeply contextualized study of single a candidate's practice) can eliminate all "error"or bad decisions. Thus, assessment developers and policy makers must decide how much (and what kind of) "error"they are willing to tolerate in light of the consequences to the candidate and to the larger system. That decision, however, will be better informed if our efforts to improve the operation of assessment systems do not obscure these problems and if the on-going research agenda undertakes to better illuminate the nature of the interpretive problem we have described here.

Large scale assessment systems must cope with the ambiguity that is on-goingly present to some degree in all efforts to understand human action. If we are to feasibly evaluate large numbers of cases, interpretive practices must be routinized, and this effort at routinization may always generate a tension amongefficiency, reliability, and validity. These challenges have led developers to processes that carefully control what raters can see and how they can interpret it. In fact, as we have repeatedly noted, all testing practices (not just

those associated with raters) place artificial constraints on how interpreters work. To one extent or another this is probably unavoidable and thus not, in and of itself, a criticism. Instead, this paper argues that the routine indicators of technical quality generally employed in current large scale assessment systems do not sufficiently illuminate the challenges of ambiguous portfolios, and that routine practices aimed at improving the quality of such systems may actually make these problems harder to find. The challenge we pose for assessment developers, then, is threefold. First, we encourage them to develop better tools for revealing the kind of challenging portfolios we have discussed here. Second, we argue that it is imperative that better ways are found for illuminating the consequences of the decisions developers make, especially around efforts to improve reader consistency. Finally, we recommend the exploration of new strategies for engaging with problematic portfolios for which "reasonable" decisions cannot effectively or ethically be made under current practices.

Notes

- 1. Authors' note: This research was supported by a grant from the Spencer Foundation. We gratefully acknowledge their support. We are also grateful to the Interstate New Teacher Assessment and Support Consortium (INTASC) and to the Connecticut State Department of Education (CSDE) for their generosity in providing the materials and opportunities that made our analysis possible. Both authors are members of INTASC'S technical advisory committee (TAC). We gratefully acknowledge comments on an earlier draft from INTASC staff and TAC members: Mary Diez, Jim Gee, Anne Gere, Bob Linn, Jean Miller, David Paradise, Ray Pecheone, and Diana Pullin. Ray Pecheone has generously read multiple drafts. 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. The role of teaching performance in licensure decisions was highlighted in a recent report of the National Research Council's committee on Assessment and Teacher Quality. The authors of the report concluded that "paper and pencil tests provide only some of the information needed to evaluate the competencies of teacher candidates" and called for "research and development of broad based indicators of teaching competence," including "assessments of teaching performance in the classroom" (NRC, 2001, p. 172). 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.
- 3. For Gadamer (1975, 1987), the hermeneutic circle had two important aspects: (a) the dialectic between the parts of a text and the whole and (b) the dialectic between the reader's preconceptions and the text.
- 4. Recent work by Peng and Nesbitt (1999) indicated that this search for coherence and the attempt to eliminate contradictions from analysis may be a somewhat western phenomenon. Peng and Nisbett compared the responses of a group of white American and Chinese college students to contradictory proverbs and propositions. They found that Chinese participants were much more comfortable with ambiguity and

contradiction. “Furthermore,” they noted, “when two apparently contradictory propositions were presented, American participants polarized their views, and Chinese participants were moderately accepting of both propositions” (p. 741).

- 5. There may also be issues arising from the requirement that the “answer” fall in one of two starkly different categories that may not capture the complexity of human motivation and situatedness. In fact, “suicide” implies clear intention which may, in fact, not have existed. Our analysis of portfolio reader processes indicates that the same challenges arise for them as well. We examine this issue in more detail in Moss, Schutz, Haniford, Coggshall, and Miller (in preparation).
- 6. The use of guiding questions that integrate standards into dimensions directed at a particular teaching performance to produce an interpretive summary, was developed by Tony Petrosky, Ginette Delandshere, Steve Koziol, Penny Pence, Ray Pecheone, and Bill Thompson in their leadership of one of the two first 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).
- 7. Although they are not an examples of ethnomethodology, we have examined aspects of the readers’processes over time in earlier work (Moss, Schutz, and Collins, 1998).
- 8. It is also important to note that in these particular cases while the two readers within each pair initially evaluated the portfolio individually, these within-pair readings were not independent: readers worked in the same room at the same time, watching the videotape together, and making comments occasionally to one another as they worked.
- 9. Quotes were edited for sense and grammar. The tense of statements was sometimes altered to fit with the paper.
- 10. Based on interviews, we have roughly estimated and depicted the range of potential scores each reader reported they considered.
- 11. In fact, Gibbs (1993) reporteda study that showed that “people find it easier to understand written language if it is assumed to have been composed by intentional agents (i.e. people) rather than by computer programs which lack intentional agency. . . . Readers presume that poets [for example] have specific communicative intentions in designing their utterances, an assumption that does not hold for unintelligent computer programs. . . . These data testify to the powerful role of authorial intentions in people’s understanding of isolated, written expressions” (pp. 190-191). In this article, Gibbs also arguedthat this tendency to impute intentions to others is a cross-cultural phenomenon, although it may appear in very different ways in different cultural contexts.
- 12. The dialogue between readers for Pair 3 of Case Study #2 was not recorded for some reason (the tape was blank), so our evidence is based only on their written notes and individual interviews. One of the readers was also different. These did not seem to raise significant issues for the case study, in total, but the lack of one tape did make it more difficult to discern differences on selected issues.
- 13. Sandra, from Pair 1, and Phil, from Pair 3, did mention subtly different score possibilities (Sandra "one plus," and Phil, "one of those . . . portfolios" that is on the "line" between a one and two) in their interviews

around the YAN portfolio, but both also indicated that they understood theseopinions were grounded in interpretations of the portfolio that were not strictly well supported by the evidence.

- 14. According to the technical manual, aresponse makes “a good training sample if a final score can be agreed on and the scoring issues presented by the response are worthy of training time” (ETS, 98, p. 60). Trainers are told during discussion to “announce the true score,” elicit “misconceptions” but “not [to] let errant assessors dominate the discussion with arguments for why your score is wrong,” and “to reiterate the true score” (pp. A33-34). Assessors are encouraged to look for the “preponderance of evidence” across “unevenness.” Unscorable portfolios are described as “weird stuff: If in live scoring, you find a case that is so incomplete or weird that you can’t score it, call the trainer to discuss it. The cases you will be seeing as training samples have all been pre-selected, so this will not be a problem for now.” [We have selected these phrases to document our assertion that assessors are not encouraged to illuminate ambiguous cases. Doing this presents a lopsided picture of The National Board’s scoring practices which are the most explicit and thoughtful we have seen within the bounds of a conventional assessment program.]
- 15. One of Myford and Mislevy’s readers, speaking about a portfolio exhibit that included contextual information, commented that it was easier to score without knowledge of the context.

References

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

Arminen, I. (1999). Conversation analysis: A quest for order in social interaction and language use. *Acta Sociologica, 42*, 251-257.

Bruner, J. (1986). *Actual minds, possible worlds*. Cambridge: Harvard University Press.

Bruner, J. (1990). *Acts of Meaning*. Cambridge: Harvard University Press.

Collins, K. C., Moss, P. A, & Schutz, A. (1997). *INTASC candidate interviews: Final summary report to INTASC.* Unpublished manuscript, University of Michigan.

Coulon, A. (1995). *Ethnomethodology* (J. Coulon & J. Katz, Trans.). Thousand Oaks: SAGE.

Davies, B. (2000). Grice’s cooperative principle: Getting the meaning across. In D. Nelson & P. Foulkes (Eds.), *Leeds Working Papers in Linguistics 8*, 1-26.

Davies, B. & Harre, R. (1990). Positioning: The discursive production of selves. *Journal for the Theory of Social Behavior, 20*(1), 43-63.

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

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. (2001). Monitoring raters in performance assessments. In Tindal & Haladyna (Eds.), *Large scale assessment programs for all students* (pp. 261-287). Mahway, NJ: Lawrence 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.

Freedman, S.W. & Calfee, R.C. (1983). Holistic assessment of writing: Experimental design and cognitive theory. In P. Mosenthal, L. Taymor & S. A. Walmsley (Eds.), *Research on writing: Principles and method (*pp.75-98). New York: Longman.

Gadamer, H. G.. (1994/1975). *Truth and method.* New York: Seabury.

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

Garfinkel, H. (1967). *Studies in ethnomethodology*. Englewood Cliffs: Prentice Hall.

Garfinkel, H. (2002). *Ethnomethodology's program: Working out Durkheim's aphorism* (A. W. Rawls, Ed.). Boulder: Roman & Littlefield.

Gee, J. P. (1990). *Social linguistics and literacies*. London: Taylor and Francis.

Gee, J. P. (1999). *An introduction to discourse analysis theory and method*. London: Routledge.

Gibbs, R. W. (1993). The intentionalist controversy and cognitive science. *Philosophical Psychology*, *6*(2), 181-206.

Goodwin, C. & Goodwin, M. H. (1992). Assessments and the construction of context. In Duranti, A. & Goodwin, C. (Eds.), *Rethinking context: Language as an interactive phenomenon* (pp. 147-190). Cambridge: Cambridge University Press.

Habermas, J. (1993). *Justification and application* (C. P. Cronin, Trans.). Cambridge: MIT Press.

Habermas, J. (1996). *Between facts and norms: Contributions to a discourse theory of law and democracy* (W. Rehg, Trans.). Cambridge: MIT Press.

Haertel, E. H. (1991). New forms of teacher assessment. *Review of Research in Education, 17*, 3-29.

Heller, J.I., Sheingold, K. & Myford, C.M. (1998). Reasoning about evidence in portfolios: Cognitive foundations for valid and reliable assessment. *Educational Assessment, 5*(1), 5-40.

Heritage, J. (1984). *Garfinkel and ethnomethodology*. Albany: SUNY Press.

Hill, C. & Larsen, E. (2000). *Children and reading tests*. Stamford, CT: Ablex.

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

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

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.

Kroeber, K. (1992). *Rereading/retelling: The fate of storytelling in modern times*. New Brunswick: Rutgers University Press.

Madaus, G. F. (1990). Legal and professional issues in teacher certification testing: A psychometric snark hunt. In J. V. Mitchell, Jr., S. L. Wise, & B. S. Plake (Eds.), *Assessment of teaching* (pp. 209-261). Hillsdale, NJ: Erlbaum.

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

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

Mink, L. O. (1987). *Historical understanding*. Ithaca: Cornell University Press.

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., Beck, J. S., Ebbs, C., Herter, R., Matson, B., Muchmore, J., Steele, D., & Taylor [Clark], C. (1992). Portfolios, accountability, and an interpretive approach to validity. *Educational Measurement. Issues and Practice*, *3*(11), 12-21.

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., Coggshall J., & Miller, R. (in preparation). High stakes assessment as ethical decision making. University of Michigan.

Moss, P.A., Sutherland, L.M., Haniford, L, Miller, R., Johnson, D., Geist, P.K., Koziol, S.M., Star, J., & Pecheone, R.L. (in press). Interrogating the generalizability of portfolio assessments of beginning teachers: A Qualitative Study. *Educational Policy Analysis Archives.*

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., & Dowling, N. M. (1993). Addressing reliability problems in the portfolio assessment of college writing. *Educational Assessment, 1*(1), 53-70.

Peng, K. & Nesbitt, R. E. (1999). Culture, dialectics, and reasoning about contradiction. *American Psychologist, 54*(9), 741-755.

Pinker, S. (2002). *The blank slate: The modern denial of human nature*. New York: Viking.

Pollner, M. (1991). Left of ethnomethodology: The rise and decline of radical reflexivity. *American Sociological Review*, *56*(3), 370-81.

Richards, R. J. (2002). "'The blank slate': The evolutionary war." *New York Times*, October 13. www.nytimes.c om, accessed 10/17/02.

Ruddell, R. B., & Unrau, N. J. (1994). Reading as a meaning-construction process: The reader, the text, and the teacher. In R. B. Ruddell, M. R. Ruddell, & H. Singer (Eds.), *Theoretical models and processes of reading* (4th ed.) (pp. 996-1056). Newark, DE: IRA.

Schegloff, E. A. (1999). 'Shegloff's texts' as 'Billig's data': A critical reply. *Discourse and Society, 10*(4), 558-572.

Smagorinsky, P. (2001). If meaning is constructed, what is it made from? Toward a cultural theory of reading. *Review of Educational Research, 71*(1), 133-169.

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., & 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.

Wolfe, E. Q. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing, 4*(1), 83-106.

About the Authors

Aaron Schutz
Associate Professor
Educational Policy and Community Studies
University of Wisconsin-Milwaukee
P. O. Box 413
Milwaukee, WI 53201
Voice: 414-229-4150
Fax: 414-229-3700
E-mail: schutz@uwm.edu

Pamela A. Moss
Associate Professor
4220 School of Education
University of Michigan
Ann Arbor, MI 48109-1259
Voice: 734-647-2461
Fax: 734-936-1606
E-mail: pamoss@umich.edu

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.

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

- Alejandra Birgin
  Ministerio de Educación, Argentina
  Email: abirgin@me.gov.ar
- Mónica Pini
  Universidad Nacional de San Martin, Argentina
  Email: mopinos@hotmail.com,
- Mariano Narodowski
  Universidad Torcuato Di Tella, Argentina
  Email:
- Daniel Suarez
  Laboratorio de Politicas Publicas-Universidad de Buenos Aires, Argentina
  Email: daniel@lpp-buenosaires.net
- Marcela Mollis (1998—2003)
  Universidad de Buenos Aires

Brasil

- Gaudêncio Frigotto
  Professor da Faculdade de Educação e do Programa de Pós-Graduação em Educação da Universidade Federal Fluminense, 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

- Antonio Teodoro
  Director da Licenciatura de Ciências da Educação e do Mestrado Universidade Lusófona de Humanidades e Tecnologias, Lisboa, Portugal
  Email: a.teodoro@netvisao.pt

USA

- Pia Lindquist Wong
  California State University, Sacramento, California
  Email: wongp@csus.edu
- Nelly P. Stromquist
  University of Southern California, Los Angeles, California
  Email: nellystromquist@juno.com
- Diana Rhoten
  Social Science Research Council, New York, New York
  Email: rhoten@ssrc.org
- Daniel C. Levy
  University at Albany, SUNY, Albany, New York
  Email: Dlevy@uamail.albany.edu
- Ursula Casanova
  Arizona State University, Tempe, Arizona
  Email: casanova@asu.edu
- Erwin Epstein
  Loyola University, Chicago, Illinois
  Email: eepstei@wpo.it.luc.edu
- Carlos A. Torres
  University of California, Los Angeles
  Email: torres@gseisucla.edu
- Josué González (1998—2003)
  Arizona State University, Tempe, Arizona

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