CHAPTER 1.

REFRAMING RELIABILITY FOR WRITING ASSESSMENT

Peggy O'Neill
Loyola University Maryland

This essay provides an overview of the research and scholarship on reliability in college writing assessment from the author's perspective as a composition and rhetoric scholar. It argues for reframing reliability by drawing on traditions from fields of college composition and educational measurement with the goal of developing a more productive discussion about reliability as we work toward a unified field of writing assessment. In making this argument, the author uses the concept of framing to argue that writing assessment scholars should develop a shared understanding of reliability. The shared understanding begins with the values—such as accuracy, consistency, fairness, responsibility, and meaningfulness—that we have in common with others, including psychometricians and measurement specialists, instead of focusing on the methods. Traditionally, reliability has been framed by statistical methods and calculations associated with positivist science although psychometric theory has moved beyond this perspective. Over time, the author argues, if we can shift the frame associated with reliability, we can develop methods to support assessments that lead to improvement of teaching and learning.

Writing an essay about reliability and writing assessment presents several challenges. One comes from determining what we mean by writing assessment because as a field it encompasses teachers and researchers in K-12 education as well as higher education. Some of these professionals are trained in educational measurement, but many others are trained primarily as literacy educators. The field also includes test developers employed by testing companies, some of whom may provide testing services for institutions, and government employees, typically in departments of education, who work on assessments such as NAEP or others. Another challenge concerns the very concept of reliability, which is deeply embedded in statistical theories and methods. Many educators who teach

DOI: https://doi.org/10.37514/PER-B.2024.2166.2.01

writing and work in college writing assessment have been educated primarily in humanities departments and are immersed in the subject of literacy education; they are not psychometricians and are not experts in statistical theories and methods, which seem to dominate approaches to reliability. Because of these challenges, college writing assessment practitioners often side-step reliability to some extent. They report instead, for example, a co-efficient about rater agreement or percentages of samples needed to be scored by three or more readers, but do not delve into the complexity of the issues associated with issues such as calculating coefficients. Yet, reliability is an important component of writing assessment that needs to be considered not just in its own right but also as part of the validation process because it addresses consistency and generalizability, among other values.

As writing assessment practitioners and scholars, we need to grapple with the challenges associated with reliability by examining how it has been used in writing assessment scholarship, especially within the college composition community, and how we can reframe it so that it both engages with what writing teachers value and contributes appropriately to validation efforts. With the interest in large-scale assessments (including writing assessment) and higher education increasing, college writing faculty will need to address several issues, many of which are related to reliability as well as validity (Adler-Kassner & O'Neill, 2010). One such issue is automated scoring, which is generally critiqued by college composition professionals (Herrington & Moran, 2001; Ericsson & Haswell, 2006) but which has found more support in the psychometric community (Williamson 2003). According to Williamson (2003):

> Two things are certain. One, automated scoring programs can replicate scores for a particular reading of student writing, and this technology is reliable, efficient, fast, and cheap. Two, automated scoring has been and will continue to be used in various large-scale assessments of student writing. (p. 256)

Carl Whithaus (2005) also acknowledged the role of automated scoring in large-scale testing and encouraged writing instructors not only to accept automated evaluation systems but also to integrate them (as well as other technologies) into their teaching (p. 13). Williamson, who doesn't go as far as Whithaus in supporting the use of automated evaluation, argued for a "productive alliance" between those in educational measurement and those invested in teaching writing (p. 101). To develop this kind of relationship, college writing instructors and program administrators need to "examine the research methodology or social sciences as it impinges on assessment" and to "explore the potential for collaborative research, not just within a social science or humanistic tradition"

(Williamson 2003, p. 101). Williamson (2003) identifies validity as a focal point of this research (p. 101), and I would add that we especially need to attend to reliability not only because of its contributions to validity but also because of the role it has played historically in writing assessment. If we can engage in these discussions, we may be able to begin reframing reliability by first developing a better understanding of reliability and then becoming full partners in the discussions—and development—of writing assessment that extend beyond our programs and institutions.

In what follows, I provide an argument for reframing reliability in writing assessment for those who come from the field of college composition as well as those whose approach is grounded in educational measurement. This analysis comes from my perspective as a composition and rhetoric scholar, but my goal is to begin a more productive discussion about reliability as we work toward a unified field of writing assessment (Huot, 2002).

THE CONCEPT OF FRAMING

In thinking about reframing reliability, I begin with the concept of framing in general. While there are many theories associated with framing, the basic idea is that we view ideas, experiences, and events through frames (akin to what Kenneth Burke called "terministic screens" and what Thomas Kuhn identified as "paradigms"). These frames usually operate at an unconscious level and are constructed by society. Members of a particular culture are conditioned to make certain connections and to understand new information through a particular frame or lens. Frames are stronger when they connect to stories shaped by the same frame (Hertog & McLeod, 2001). Communication and cultural theorists, such as Stuart Hall (1983), have explained that the media play a dominant role in this process through cultural conditioning, which is established by the boundaries media set around the stories that they cover. These boundaries, which determine what is and is not covered, create tacit connections and connotations for members of that culture. For example, current debates about education policy and funding are framed by the concept that education should prepare students for college or careers. This has become so ubiquitous in the media that it is hard to articulate other purposes of education, such as civic engagement. This perspective links to other stories about education, such as the often repeated story about US students lagging behind other students, which is reinforced by, among other things, the "Race to the Top" initiatives supported by the US Department of Education. In today's culture with non-stop access to news through 24-hour cable channels and the Internet, mass media is an especially powerful means of creating frames.

Cognitive linguists acknowledge the importance of culture and media in creating frames, but they take it deeper. Framing, according to cognitive linguists, is about how our minds work to make meaning from language and images. George Lakoff (2006) explained that framing is "a conceptual structure used in thinking" and that every word evokes a frame. In fact, words have no meaning outside of frames, which fit together to form systems. The frames are reinforced the more they are evoked: "Every frame is realized in the brain by neural circuitry. Every time a neural circuit is activated, it is strengthened" (Lakoff, 2006, n.p.). In other words, frames are connected to the way our brains are wired. Every word evokes a frame, and words defined within a frame, evoke the frame. The example Lakoff uses to explain this concept is "Sam picked up the peanut with his trunk," which evokes the frame of elephant because we understand trunk in this sentence within that frame. Even negating a frame evokes it as in Lakoff's example, "Don't think of an elephant!" which is impossible to carry out because as soon as elephant is mentioned, we think of it. Every time a frame is evoked—whether negatively or positively, whether directly or indirectly—it is strengthened. So in arguing against a frame, we are actually reinforcing it.

Frames are so powerful that they shape the way we understand facts. Facts, explained Lakoff (2002), are understood within our frames so that people with different worldviews understand and process the facts differently. In other words, as Lakoff (2002) argued, it isn't a matter of just getting the most accurate information into the debate, it is critical that that information is framed in ways that make sense to the audience. Other ways of reasoning, including framing and categorization, creates "huge variability in normal, everyday human reasoning" (Lakoff, 2002, p. 373). What one person sees as clear, rational commonsense can be understood in completely different ways by others depending on the individuals' frames, which makes communication more difficult. Many academics and researchers may experience this kind of communication disjunction when trying to discuss issues related to their scholarly work with a non-expert. Sometimes the difficulties are simply related to terminology, for example, the term "grammar" is often used by non-experts to discuss the teaching of writing to cover a wide range of issues to address in teaching writing from mechanics and punctuation to style, organization, use of evidence and a myriad of other aspects. In this type of situation, further discussion and probing can usually clear up the confusion. However, communication problems can also be rooted in different frames.

How individuals frame a concept such as "teaching writing" will depend on their own experiences, education, expertise and values as well as how it has been depicted in the culture. This understanding, furthermore, may or may not align with what a particular person means when she uses the phrase "teaching writing." An individual's frame will be reinforced every time it is evoked. So, for example,

when the Common Core State Standards Initiative identifies writing as one of the key areas for "college and career readiness," readers will understand this section through the frame they have already have about writing. Parents, teachers, policymakers and assessment experts may not all share the same framework and so they may understand the standards differently. The authors of the standards try to mitigate this situation by providing preliminary material that defines terms, explains situations, and even articulates what is not covered by the standards. However helpful that information is, it can also act as a away to reinforce what the reader already thinks and believes because many of the associations and assumptions work at unconscious levels. If we consider writing assessment, then, the same theory applies. What seems practical and rational to writing teachers may seem completely unreasonable or just wrong to policymakers or psychometricians, who may approach the activity through completely different frames—different values, experiences, assumptions, and world views. Specific technical terms associated with assessment, such as validity and reliability, will also be understood differently depending on the frame surrounding them.

If we want to change a concept or redefine a concept, we need to consider the frame that surrounds it and how that influences the way a term is understood. Trying to change the term without taking into account the bigger picture will not be successful, in Lakoff's view, because much of what is evoked happens automatically and unconsciously. Making visible the dominant associations and assumptions so we can see the frame that currently in place is the first step in trying to reframe writing assessment in general and reliability in particular.

TRADITIONAL FRAMING OF WRITING ASSESSMENT

For the last hundred years, reliability has been the dominant frame surrounding writing assessment, pre-occupying scholars and test developers (Huot & Neal, 2006; Williamson, 1993; Huot, 2002; Elliot, 2005; O'Neill, Moore, & Huot, 2009). Although reliability, as a psychometric concept, encompasses a broad range of concerns, in writing assessment this quest has focused primarily on scoring, specifically getting scorers to agree at an acceptable rate, which is referred to as interrater reliability. As Huot and Neal (2006) concluded in their techno-history of writing assessment: "Throughout the history of writing assessment and whether we refer to technologies like the indirect tests of grammar usage and mechanics, the use of rubrics and rater training, or the machine-scoring of student writing, we are basically referring to technological solutions to the problem of scoring consistency" (pp. 418-19). For example, the College Entrance Examination Board (CEEB) ostensibly abandoned essay exams in 1941 as part of the war effort to streamline student matriculation for potential armed forces

recruits. In truth, the CEEB had been piloting the SAT for scholarship students who needed to apply earlier and had found that the reliability and efficiency of the SATs to be much superior to that of the essay examination. The development of holistic scoring procedures in the 1960s, done by Educational Testing Service researchers Godschalk, Swineford, and Coffman (1966), revitalized essay testing because it provided a reliable way to score essays.

By the 1980s, holistically scored essays enjoyed widespread use for a variety of writing assessments across educational levels, but especially in college. Edward White (1993) claimed: '[W]hen a university or college opens discussion of the measurement of writing ability these days, the point of departure is usually a holistically scored essay test" (p. 89). The holistic scoring of essay exams depended upon standardization of procedures for the test administration, of the tasks and topics, and of the scoring. The holistic scoring sessions became, according to White (1993), not just a method for scoring but also a means of professional development as readers discussed anchor papers and practiced scoring samples to internalize the scoring rubric so they could apply it in a consistent way. These scoring sessions also required careful record keeping and checks for agreement between two independent raters.

While White focused on the benefits of holistic scoring both in terms of professional development and achieving acceptable reliability rates, Cherry and Meyer (1993) critiqued the way reliability has been handled in writing assessment. They explained that reliability "refers to how consistently a test measures whatever it measures" (p. 110). The consistency of a measurement, Cherry and Meyer (1993) explained, can come from the test design and administration, the students, or the scoring. For essay testing, particular sources of error may include the prompt—which may not produce reliable results—as well as the administration and scoring of the essays. After reviewing research in direct writing assessment from Starch and Elliot's 1912 article, "Reliability of the grading of high school work in English," through several pieces in the mid to late 1980s, Cherry and Meyer (1993) concluded that there have been four serious problems with reliability as reported in writing research and evaluation (p. 116).

First, according to Cherry and Meyer (1993), reliability discussions (with a few notable exceptions) have been limited to interrater reliability although there are many other aspects of reliability that need to be considered. For example, if students' performances are not accurate in terms of their writing abilities because of the prompt design, then results are not reliable no matter how consistently raters apply the rubric and how much they agree with each other (Hoetker, 1982).

Second, there has been confusion over reliability and validity in influential studies of writing assessment. Cherry and Meyer (1993) critiqued Godshalk,

Swineford, and Coffman (1966) because they identified differences in results across topics as a reliability issue when in fact these differences are about validity. Variation across topics/prompts, they explain, can be a validity issue because the underlying construct being tapped is different if the writing tasks are different.

Third, there has been a lack of agreement on appropriate statistical methods for determining interrater reliability. Cherry and Meyer (1993) reported that at least eight different methods had been used in computing and reporting interrater reliability statistics and that many studies never even explained how they calculated the reliability co-efficient (p.119). However, the variable ways for calculating the interrater reliability co-efficient can yield drastically different results. For example, using a straight percentage of agreement between raters, the Pearson correlation coefficient or Cronbach's alpha to calculate the interrater reliability will produce different statistics for the same data.

Finally, various procedures used in holistic scoring sessions directly affected the reliability statistics. Cherry and Meyer (1993) explained that sometimes reliability rates are based on "practice sessions," not live scoring, which can artificially inflate the interrater reliability statistic. Other problems come from the practice of "resolving" differences between two raters by using a third rater. In fact, Cherry and Meyer (1993) recommended discontinuing the practice of resolving differences all together (p. 122) because "interrater reliability formulas are quite sensitive to the manipulation of data" through these methods even when a low percentage of scores are affected (p. 123). Haertal (2006), in discussing reliability and ratings of products, echoed their concern: "It must be emphasized that when adjudication is used, assumptions for many statistical models are violated" (p. 102).

Stemler (2004) argues that interrater reliability needs to be unpacked. He contends "the widespread practice of describing interrater reliability as a single, unitary concept is at best imprecise, and at worst potentially misleading" (p. 2). He identifies three categories of interrater reliability—consensus estimates, consistency estimates, and measurement estimates—and details the assumptions, interpretations, advantages, and disadvantages of each (p. 2). The statistical methods for determining the different types of interrater reliability also vary, and Stemler (2004) reviews these as well. Although Stemler (2004) is not limiting his focus to writing and literacy assessments, he seems to agree with Cherry and Meyer (1993) that researchers do not address the nuances of interrater reliability enough.

While Cherry and Meyer (1993) articulated several problems with reliability reported in writing assessment research, Hayes and Hatch (1999) focused on problems with reliability in literacy research in general, including rating of student work whether for a testing or research purposes. Hayes and Hatch (1999)

also critiqued the method of calculating and reporting reliability found in the literature, especially on more recent studies. They argued that interrater reliability rates should be determined by statistical correlations and not the percentage of agreement between the two independent raters. Hayes and Hatch (1999) explained that reliability calculated using a statistical correlation formula takes into account the role of chance in the agreement rate while the percentage method doesn't. Depending on the scoring scale and the distribution of scores, chance can account for a significant portion of agreement. For example, the fewer score points on the rating scale, the greater the influence chance has on the agreement rate; or, the more scores tend to cluster around certain scores, the more influence chance has on the reliability measure.

Like Cherry and Meyer (1993), Hayes and Hatch (1999) noted that different methods for calculating reliability lead to different results, yet they also found many researchers did not report the method for calculating reliability correlations. Both Cherry and Meyer (1993) and Hayes and Hatch (1999) also agreed that when researchers do not fully disclose how they determined reliability estimates, it is difficult for readers to determine if the method is appropriate, to compare reliability across studies, and to avoid confusion. Hayes and Hatch (1999) concluded their essay with an acknowledgment that other methods exist for measuring reliability, including generalizability measures, than those they address although they don't discuss them.

Both Cherry and Meyer (1993) and Hayes and Hatch (1999) framed reliability in writing assessment using classical test theory. Shale (1996), however, advocated using generalizability theory instead, arguing that it is more appropriate for addressing the issues associated with reliability in writing assessment because it can address the multiple sources of error that can arise in a writing assessment. In most writing assessments, Shale (1996) contended, reliability is vague because it is only considered within the classical test theory, which was developed for multiple-choice testing: "Considerable ambiguity arises because the full sense of reliability as understood within the context of multiple-choice testing does not transfer well to the world of essay testing" (p. 77). Shale explained that the consequences of considering reliability only in terms of classical test theory has resulted in a "fixation on marker disagreement" which has led to a distortions and limitations in writing assessment practices (p. 78). Shale (1996) as with Cherry and Meyer (1993) and Hayes and Hatch (1999), also noted the paucity of rigorous inquiry into reliability in writing assessment scholarship. Reliability, how we should approach it, and what we mean by the term is still an issue in college writing assessment as I discuss in more detail later.

While concerns about reliability of essay exams preoccupied writing assessment scholars for a long time and, in effect framed writing assessment, the

validity of essay testing was not seriously challenged because essay testing required students to write instead of completing multiple-choice items about language conventions and grammar. White (1993) articulated the assumptions that supported holistic scoring of essay exams: "It is a direct measure of writing, measuring the real thing, and hence is more valid than indirect measures" such as fill in the bubbles multiple choice exams and editing tests (p. 90). By the 1990s, however, writing assessment scholars (as well as measurement theorists) began to turn their attention to validity arguing that a portfolio of writing was preferable to a single-sample, timed impromptu essay (Elbow and Belanoff, 1986).

The shift to validity began to take the focus away from reliability as a purely statistical concept and to frame it as part of a validity argument, which addresses both theoretical and quantitative, statistical evidence (Messick, 1989). Camp (1993) addressed the tension between classical test theory and emerging theories of writing and literacy. Camp (1993) argued: "Very likely we are seeing the signs of a growing incompatibility between our views of writing and the constraints necessary to satisfy the requirements of traditional psychometrics—in particular, of reliability and validity narrowly defined" (p. 52). Camp (1993) explored this tension, identifying some of the key factors that may need to be addressed to develop writing assessments that take into account what we know about writing as well as the principles of fairness, equity, and generalizability—concepts, she explained, that are associated with reliability. The challenge, according to Camp, has been to apply these principles in ways that lead far beyond the narrow focus on score reliability and constricted definitions of validity that have characterized earlier discussions of writing assessment (p. 68). At the time of Camp's essay, portfolios (like other performance assessments) were growing in popularity and Camp concluded with a brief discussion of some portfolio projects.

Since the early 1990s, the popularity of portfolios in college writing programs has continued to spread for both teaching and assessment (although essay testing also remained popular). Although writing portfolios seemed to be a substantive departure from impromptu essay testing, the discussion of reliability, however, did not changed very much. The focus was still narrowly on interrater reliability. As White (1993) looked to the future of portfolios, he identified reliability of portfolio scoring as the major issue to deal with, which in effect, continued to frame writing assessment in terms of reliability. At that time, he recommended adapting many of the same procedures for portfolios that were used for holistic scoring of essays: "At a minimum, each portfolio should receive two independent scores, and reliability data should be recorded. While reliability should not become the obsession for portfolio evaluation that it became for essay testing, portfolios cannot become a serious means of measurement without demonstrable reliability" (p. 105).

Compositionists in college writing programs, following Elbow and Belanoff (1986), developed an assortment of writing portfolio assessments for placement into first-year writing (Willard-Traub, Decker, Reed, & Johnson, 1999; Daiker, Sommers, & Stygall, 1996; Borrowman, 1999; Lowe & Huot, 1997; Hester et al., 2007) and proficiency (Roemer, Shultz, & Durst, 1991; Nelson, 1999; Haswell, 2001). Many of these portfolio assessments adapted holistic scoring methods used for essay exams to portfolios, reporting the interrater reliability and in many cases, doing so in ways that are problematic according to Cherry and Meyers (1993), Shale (1996), and Hayes and Hatch (1999). For example, Borrowman (1999), reporting the reliability of placement portfolio system at the University of Arizona, presented the reliability co-efficient for the program but did not explain how the figure was calculated. He did, however, devote three pages to discussing reliability and how the high interrater reliability is achieved: "the physical conditions in which the scoring of portfolios takes place and the generation of the scoring rubric" (12). Borrowman (1999) addressed the tension between reliability and validity but he only considered interrater reliability in his discussion, which is a very limited understanding of reliability.

RECONSIDERING THE TRADITIONAL FRAMEWORK

While White (1993) was correct that reliability is a critical issue to address, his assumption that the same methods associated with holistic scoring are the minimum requirements for portfolio assessment demonstrates how writing assessment practitioners and scholars often have a limited reliability as a theoretical construct. It also illustrates how the narrow psychometric frame continued to dominate many of the discussions of reliability in college composition. Yet, in spite of the focus on validity, the critique of traditional treatment of reliability in writing assessment, and discussions about scoring and reliability, many college writing assessment programs still failed to address the reliability issues that Cherry and Meyer (1993) identified in the literature associated with essay exams. While a few writing assessment scholars began pushing against reliability (Smith, 1992, 1993; Haswell & Wyche, 1996; Broad, 1994; Lowe & Huot, 1997), as a field we didn't grapple with it too directly. So when we encountered Moss's (1994) question, "Can we have validity without reliability?" we seemed to respond with an enthusiastic "Yes!" Reliability, however, is an important theoretical construct, and can't be dismissed or ignored. Mislevy (2004), as part of a special section of the *Journal of Educational and Behavioral Statistics*, responded to Moss's (1997) question as well as other commentaries on it, explaining that reliability in psychometrics encompasses a wide range of issues so that "a measure wholly unreliable in the more fundamental sense would consist only of error and could not support valid inferences" (p. 1).

We need to explore the concept more fully, considering it in light of what we know about writing and learning to write, as well as psychometric theory, because as Camp (1993) said, the principles that inform reliability are important.

Moss (1994), in fact, didn't reject reliability outright. Rather she encouraged assessment researchers and practitioners to explore it as a theoretical construct in light of validity. She explained that "less standardized forms of assessment . . . [such as portfolios] present serious problems for reliability, in terms of generalizability across readers and tasks as across other facets of measurement" (p. 6). Though carefully trained readers can achieve acceptable rates of reliability, Moss (1994), an educational measurement theorist, argued that with "portfolios, where tasks may vary substantially from student to student, and where multiple tasks may be evaluated simultaneously, inter-reader reliability may drop below acceptable levels for consequential decisions about individuals or programs" (p. 6). Moss concluded that "although growing attention to the consequences of assessment use in validity research provides theoretical support for the move toward less standardized assessment, continued reliance on reliability, defined as quantification of consistency among independent observations, requires a significant level of standardization," (p. 6). However, these less standardized forms of assessment are often preferable "because certain intellectual activities" cannot be documented through standardized assessments (p. 6).

Moss (1994) suggested that in educational assessment, we look beyond psychometric theories and practices in cases where acceptable reliability rates are difficult or impossible to achieve. She challenged the assessment community to consider its definitions of reliability—and here we in writing assessment need to remember that reliability is more than a quantification of consistency among independent observations. Moss recommended a hermeneutic approach because as a philosophical tradition, it values a "holistic and integrative approach to interpretation of human phenomena" (p. 7). After summarizing the key perspectives of hermeneutics, Moss explained how this methodology would work:

> A hermeneutic approach to assessment would involve holistic, integrative interpretations of collected performances that seek to understand the whole in light of its parts, that privilege readers who are most knowledgeable about the context in which the assessment occurs, and that ground those interpretations not only in textual and contextual evidence available, but also in a rational debate among the community of interpreters. (p. 7)

Critical features of this type of assessment include the recognition of disagreement or difference in interpretations as evaluators bring their expertise and experience to bear on the work. Positions of individual evaluators can change

as rational debate ensues, with the final decision coming out of consensus or compromise. In supporting this approach in specific situations, Moss (1994) reminded readers that reliability and objectivity are no guarantors of truth and that they can, in fact, work against "critical dialogue" and can lead "to procedures that attempt to exclude, to the extent possible, the values and contextualized knowledge of the reader and that foreclose[s] on dialogue among readers about specific performances being evaluated" (p. 9). Mislevy (2004), saw benefits in Moss's idea but also commented:

> In assessment, as in other fields, difficulties arise when novel problems appear and the usual heuristics fail. We now envisage assessments that target inferences more subtle than proficiency in a specified domain of tasks. . . . We must return to first principles to establish the credentials of this evidence . . . The hermeneutic tradition does offer insights into drawing inferences from disparate masses of evidence, and we can indeed learn much from dialectic between psychometrics and hermeneutics. (p. 2)

He advises, though, that a first step is to acquire "a deeper understanding of psychometric methods, an understanding of principles behind methods that will not be found in common wisdom, familiar testing practices, or standard textbook presentations" (p. 2).

Moss's comments about a hermeneutical approach to complex performance assessment echoed what writing assessment scholars praised about holistic scoring sessions and alternative methods for evaluating student writing (whether portfolios or essays). White (1993, 1994), who has been a stalwart supporter of holistic scoring of student writing has often expounded on the benefits associated with norming and scoring sessions. Scholars, reporting on portfolio assessments, made similar statements such as Hamp-Lyons and Condon (2000):

> Instead of focusing on scores, readers spend time bringing their reading processes into line with each other. They read and discuss samples with an eye toward developing and refining a shared sense of values and criteria for scoring. In other words, this method fosters a reading community in which reliability grows out of the readers' ability to communicate with each other, to grow closer in terms of the ways they approach samples. (p. 133)

Although Hamp-Lyons and Condon (2000) addressed reliability in this way, they still used more traditional reliability evidence to justify portfolio assessment:

"The reliability obstacle, in some local contexts, has been overcome. Miami University's reliability statistics, like Michigan's, are within the .8 range of holistic essay assessments . . ." (p.91). Their position echoed White's (1993) concerns about portfolios. However, Hamp-Lyons and Condon (2000) and others did not address the concerns about reliability articulated by Cherry and Meyers (1993), Shale (1996), and Hayes and Hatch (1999).

Other scholars pushed against the traditional holistic scoring approach designing methods that privileged those most knowledgeable about the context, that encouraged critical dialogue, and that used holistic and integrative judgments. Smith (1992, 1993) found that placement decisions for students entering college composition were more reliable with an expert reader system than when made via traditional holistic scoring procedures. In Smith's system, readers made decisions based on the most recent course they taught, either accepting or rejecting the student for the course or rejecting. Haswell and his colleagues at Washington State University (Haswell & Wyche, 1996; Haswell, 2001) developed a two-tiered expert reader system in which readers made the initial decision of whether or not a student should start in the regular composition course—the one most students take. A panel of expert readers made decisions for those students who did not fit neatly into this course. In making their decisions, the panel of readers could consult and discuss difficult cases instead of following the standardized, objective procedures associated with holistic scoring. Writing program administrators at the University of Cincinnati used a system of portfolio assessment to replace the first-year composition essay exit exam (Roemer, Shultz, & Durst, 1991; Durst, Roemer, & Shultz, 1994). The portfolio scoring system used large group "norming" sessions in conjunction with trios of writing teachers who worked independently to determine if students met the basic requirements to successfully exit the composition program. These alternative systems were still interested in reliability but not in achieving acceptable rates through the conventional approach to holistic scoring.

Others (e.g., Broad, 1994; Lowe & Huot, 1997; and Hester et al., 2007) challenged the traditional holistic scoring approach that characterized most portfolio assessments. Broad (1994) and White (1993, 1995) represented the concerns about reliability that circulated around the use of portfolios as a large-scale assessment method, but writing assessment scholars as a field still did not interrogate the concept of reliability. More recently, White (2005) noted the difficulty in reaching acceptable reliability rates that has plagued portfolio assessments and proposed a scoring method for portfolios "derived conceptually from portfolio theory, rather than essay-testing theory" (p. 583), overturning his earlier position that portfolios are basically just expanded essay tests (White, 1995). Although White (2005) seemed to be advocating a method of portfolio evaluation distinct from holistic

scoring, he describes his approach, which focuses on the reflective letter or self-assessment and clear statements of learning goals, this way:

> Now we can speak sensibly of scoring, even holistic scoring, of the reflective letter, which needs to meet certain quite specific criteria. We are back to a single document, the basic material for which holistic scoring was designed, and we can usually agree on the quality of that document, though we may disagree on the quality of the items in the portfolio that support that document. With some labor, we can come up with a scoring guide and sample portfolios at various score points, just as we can do with single essays. (p. 593)

In short, White's new method was closely aligned with the old one and was designed to streamline the portfolio scoring by focusing on a single text. Granted, he explained how the portfolio contents were used along with the writer's self-assessment, but he still framed of reliability in traditional conventional ways.

While Moss (1994) recognized that reliability standards, within the psychometric tradition, are grounded in fairness to stakeholders, she contends that from a hermeneutic perspective, reliability "can be criticized as arbitrarily authoritarian and counterproductive" (pp. 9-10). In the end, Moss did not argue for abandoning reliability but rather advocated that alternative approaches to assessment theory and practice be considered when appropriate (p.10). Her position is especially relevant for those charged with writing assessments because writing is a complex, multidimensional, contextually situated activity. Importing psychometric theory and practices, especially in terms of reliability, may undermine the very usefulness of a writing assessment's results. However, psychometric theory cannot be dismissed out of hand; instead, writing assessment scholars and practitioners need to draw on language, literacy and psychometric theories as well as other interpretive traditions to design assessments. Some scholars in college composition have done this (Smith, 1992, 1993; Haswell & Wyche, 1994; Broad, 1994, 2003; Lowe & Huot, 1997; Huot, 2002) there are still many assessment practitioners who conform to more narrow approaches, relying on an interrater reliability statistic to demonstrate reliability as we saw with Borrowman (1999).

REFRAMING RELIABILITY

Moss's (1994) argument to reconsider reliability through alternative research traditions appeals to those of us in writing assessment more comfortable with literacy studies, literary theory, and qualitative research methods. However, it

doesn't necessarily resolve some of the conflicts we experience in confronting the pre-occupation with reliability, narrowly conceived, that dominates large-scale assessments. College composition scholars Penrod (2005) and Lynne (2004) argued that psychometric concepts such as validity and reliability are not pertinent for college writing assessment because they are rooted in a positivist epistemology that is incompatible with the social constructivist approaches of writing and meaning-making that inform most of the field's work. Both Penrod (2005) and Lynne (2004) drew on qualitative research traditions. Lynne (2004), who used Guba and Lincoln among other theorists, suggested isolating college composition from educational measurement and developing our own key assessment terms. She offered "meaningfulness" and "ethics" for use instead of validity and reliability (p. 117). While both Penrod's and Lynne's critiques of validity and reliability (and psychometric practices in general) addressed some important concerns, if we attempt to reject reliability, or ignore it, we will make writing assessment more vulnerable to methods and interpretations of results that contradict what we know about literacy and writing, ultimately compromising validity. Like Lynne (2004), Huot (2002) advocated for assessments that are meaningful (p. 101) and acknowledged our responsibilities in writing assessment (p. 57-58), but he called on us to participate as full partners with educational measurement colleagues (p. 57). Psychometric theory is, after all, compatible with what writing teachers and scholars value even if these shared values are not always emphasized in practice (Huot, 2002; O'Neill, Moore, & Huot, 2009).

While Huot's (2002) discussion of validity and reliability have been acknowledged as making a significant contribution to the field of writing assessment, most writing administrators and writing faculty are not seriously engaged in theoretical discussions of assessment in general or reliability in particular. We are often too focused on practice—solving an immediate need, refining an existent assessment procedure—to engage in theoretical debates about assessment terms and principles (Gere, 1980; Faigley et al., 1986). Many people charged with college writing assessments are not composition scholars let alone writing assessment experts. We can't reject basic principles or terms, especially when a term is invested with so much cultural capital and power as reliability, without a better understanding of what reliability brings to the table and what it represents in the wider assessment community.

Since the early 1990s, we have seen an assortment of assessment models (e.g., Smith 1992, 1993; Haswell, 1994, 2001; Broad, 2003) that challenge conventional approaches to reliability; however, most of us are still confronted with demands for reliability narrowly framed or are ill-prepared for discussions about reliability. What happens, in many cases, is that those of us charged with writing assessment, who also identify as literacy teachers and researchers, have found

ourselves in discussions with testing specialists, whether in our institutions or from outside vendors, but unable to communicate clearly with them. We need to think carefully about what values reliability taps into and how they connect to the values we hold about teaching writing and learning to write. As Haertel (2006) concluded in his discussion of reliability—which is not specific to writing assessment—we need "further integration of notions of reliability with evolving conceptions of test validity" (p. 103). We need to understand, as Camp (1993) argued, the principles that reliability encompasses. We also need to think more strategically about identifying what we value and how to communicate that in ways that will be persuasive to others—policymakers, administrators, test developers. As Lakoff (2004) reminded us, language not only provides form for our values, ideas, experiences and thoughts, it helps shape them and how we understand the world around us. This often occurs unconsciously so we need to be intentional and thoughtful about how we use language to frame writing assessment and reliability or we may be undermining our own efforts.

If we think more strategically, as Lakoff (2004) recommends, about how we want to frame reliability—and writing assessment more generally—we need to consider how we use reliability and what it evokes with the educational assessment culture, especially in the field of writing assessment. As noted above, reliability has been a longstanding issue in educational measurement and in writing assessment. It is associated with quantification—measurement, scoring, statistics—and it also evokes validity.

In some sense, however, college writing assessment as a field of study seems ambivalent toward reliability. As Cherry and Meyer (1993) explained, writing assessment practitioners have not been consistent in the methods or presentation of interrater reliability although we keep using the term and providing a co-efficient. By continually referring to reliability and presenting a statistic, we have reinforced the traditional frame for evaluating an assessment. Yet, we have not established consistent methods in determining reliability or even in discussing how we are approaching it and why. As the Hamp-Lyons and Condon (2000) example above illustrates, we seem to try to have it both ways: we report a reliability statistic but what we find most valuable is the discussion and debate we have to develop the community of knowledgeable readers. Peckham (2009) illustrates some of the difficulties college composition as a field has with reliability. Writing about a pilot online placement system in the flagship journal of the Conference on College Composition and Communication, Peckham (2009) addressed scoring of placement essays and the results compared to the placement students received based on the ACT score. He acknowledged Huot's (1996) more recent critique of interrater reliability but then argued for it in terms of values (fairness), which he equated to validity. He wrote:

> Although Brian Huot argues interrater reliability is over valued ("Toward a New Theory," p. 560), I think there is some question about the fairness (and thus the validity) of an assessment if readers frequently disagree on the placements. (p. 521)

Later he explained, in part, the results of the essay scoring he conducted, noting the interrater reliability of the readers (the specific term interrater reliability isn't used). Peckham also acknowledged that the high rate of agreement could be a problem if the raters were consistently wrong but seems to dismiss this as unimportant. He wrote::

> Of course, the notions of "right" or "wrong" are highly suspect in any discussion of writing assessment. The only thing we can say with confidence is that we recommended reassignment for about 42 percent of the students. The percentages in the differences and directions of recommended reassignments over the three years suggest that our five readers, who remained generally the same from year to year, were at least ranking the essays consistently. *We picked the readers from among the best teachers in our program. . . . Unsurprisingly, our agreement rate was high—only 3 to 4 percent of the essays in the three years needed a third reading. Admittedly, reader agreement does not guarantee a valid assessment; my readers could be consistently wrong, but assessment is not a question of wrong or right*: it is about best choices, in this case to place students on the basis of their writing and a controlled scoring or on the basis of multiple-choice exam. (Peckham 2009, p. 535; emphasis added)

Peckham (2009) also addressed other aspects of reliability, such as the reliability of the test itself and connected that to scoring. Two aspects of this selection, excerpted below, are noteworthy: 1) After acknowledging the benefit of using two writing samples, he explained they use one because of "simplicity," which as a value seems to be prized more than reliability; and 2) He implied that he isn't confident in the abilities of the raters' scoring, which seems contradictory to what he said about the raters' agreement above:

> [W]e realize that for a more reliable assessment, we should require at least two essays for two reasons: first, two essays in different genres might increase "test reliability," that is, that given similar testing situations, students will achieve relatively

> similar scores on both tests (White, "Apologia," p. 41); and second, the second essay would allow us to assess the student's ability to respond to a writing task based on one of the major assignments in our second semester course. But *we decided to forgo the probable increase in test reliability for greater simplicity*. Our experience has shown us that it is difficult to train teachers to agree on the criteria and rankings of anchor papers in one genre. *When we are confident about teachers' abilities to score essays in one genre, then we will move to two essays in different genres*. We expect to expand our submissions into electronic portfolios, but that's down the road. (Peckham 2009, p. 526; emphasis added)

Peckham knows reliability is important, but he also seems to indicate that there may be some problems with it as it applies to a writing assessment. He wants the assessment to be fair, and "valid" (p 521) and he believes consistency in the scoring is important (pp. 526 and 535). But he favors simplicity over other concerns about the test. After reviewing research on the correlations between direct and indirect methods of writing assessment, Peckham (2009) concluded that "I would go with the writing simply because we are more nearly looking at what we think we are trying to assess (i.e., the direct method has more testing validity)" (p. 532). Peckham's article illustrates how as a field, there is some degree of uncertainty about how to handle all the nuances and technical components of reliability (and, by extension, validity).

The point in detailing Peckham's references to reliability is not to critique him per se but rather to illustrate the ambivalence we as a field have around reliability and the difficulty we have in addressing it. His article, after all, was published in *CCC* which "reflects the most current scholarship and theory in the field," according to its website, and uses blind peer-review. Because of its publication in *CCC*, Peckham's discussion of reliability also serves as a powerful example of our reluctance to address the concept of reliability more directly and in more theoretically informed ways. It demonstrates how a purely quantitative, statistical approach to reliability does not fit well with what we value. However, it also shows that we recognize the significance of reliability and that there are some positive, useful values that reliability supports, so we cannot dismiss it out of hand. This is what Lynne (2004) realized in her attempt to replace the terms *validity* and *reliability*. However, while we might need to consider the language—as Lynne (2004) suggested—we need to focus on what we value, what concepts are most important, and what ideas are involved when discussing reliability because frames are ultimately about values, ideas and concepts—the

language merely evokes and reinforces the frame. Therefore, we need to be more intentional and thoughtful about the language we use in discussing reliability and writing assessment.

Lakoff (2006) explains that language choice is "vital" because "language evokes frames—moral and conceptual frames" (p. 7). So far, we have allowed the psychometric practitioners (and I would also argue conservative policymakers and their constituencies) to frame reliability in ways that privilege their worldview and support their values. We need to consider ways to reframe reliability so that it evokes the values that literacy teachers hold and support in their research about teaching, learning, and language. Thinking about reliability as a concept, an issue, as well as the frame it evokes and how we can communicate more effectively about what we value, is a role that literacy educators are able to tackle because it shifts the debate away from statistical methods and technical expertise to the concept of reliability, the values it promotes, and the ways these values are communicated (Parkes, 2007).

While few writing teachers and theorists are psychometricians or experts in advanced statistics, many more are experts in language and literacy. We understand communication theory and language development. We know about teaching, learning, and students. We have strong values and beliefs—such as a belief that all children can learn, that all deserve access to quality education, that context is critical in effective writing, and that writing assessment should improve teaching and learning. Smith (1992, 1993) explored multiple aspects of reliability in a series of ongoing studies that were, in effect, a process of validating the locally-designed placement system he developed (O'Neill, 2003). His goal was to make sure that students in his program were placed in the most appropriate first-year writing course. Huot (2002) in arguing for a new theory of writing assessment that values context, local control, rhetorical principles, and accessibility considered reliability as part of the validation process.

Reframing a concept as ingrained and complex as reliability requires a commitment because frames are developed overtime, unconsciously in most cases, through repetition and reinforcement. Everyone has frames—and they are not always theoretically consistent or compatible—although people are usually not aware of them because they function at the unconscious level. To reframe an issue, Lakoff (2006) explained, we need to be strategic. With reliability, we can start by determining how it has been framed and then how we can reframe it in ways that support our beliefs about teaching and learning. One place to start is with the standard reference manuals in the field of psychometrics. To that end, below are excerpts of basic explanations of reliability from the most recent editions of two mainstream measurement reference manuals: *The Standards of Educational and Psychological Testing* (AERA, APA, & NCME, 1999) and

Educational Measurement, 4th edition (ACE, 2006). From the *Standards*, here's the opening paragraph on the section "Reliability and Errors of Measurement":

> A test, broadly defined, is a set of tasks designed to elicit or a scale to describe examinee behavior in a specified domain, or a system for collecting samples of individual's work in a particular area. Coupled with the devise is a scoring procedure that enables the examiner to quantify, evaluate, and interpret the behavior or work samples. *Reliability* refers to the consistency of such measurements when the testing procedure is repeated on a population of individuals or groups. (p. 25)

According to the glossary in the *Standards* (AERA, APA, & NMCE, 1999), reliability is "the degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be repeatable for an individual test taker" (p.180). It also includes the "degree to which scores are free of errors of measurement for a given group" (p. 180). Haertel (2006), in the fourth edition of *Educational Measurement*, opens the chapter on reliability this way:

> The concern of reliability is to quantify the precision of test scores and other measurements . . . Like test validity, test score reliability must be conceived relative to particular testing purposes and contexts. The definition, quantification, and reporting of reliability must each begin with considerations of intended test uses and interpretations. However, whereas validity is centrally concerned with the nature of the attributes tests measure, reliability is concerned solely with how the scores resulting from measurement procedure would be expected to vary across replications of that procedure. Thus reliability is conceived in more narrowly statistical terms than is validity. (p. 65)

Both of these explanations highlight the statistical, technical apparatus that typically frames reliability. In this frame, quantification and measurement are invoked. Measurement implies a finite amount of something. This epistemology is associated with objectivity that was the central to psychometrics in the early and mid-twentieth century (Williamson, 1993, 1994). However, in these excerpts, values of consistency and accuracy are also identified. Haertel (2006) even acknowledged context as a value when he notes that it "must each begin with considerations of intended test uses and interpretations" (p. 65) since these aspects of an assessment will define, in part, the particular situation. And in

fact, these are also values that were central to the development of psychometrics. Parkes (2007) argued that reliability as a concept has been conflated with its methodology and that what we need to do is remember that it is the values that are primary. Camp (1993) made a similar point. The methods to demonstrate reliability should not be more important that the values that reliability represents. Parkes (2007) explained it this way:

> The outcomes of the use of these tools—reliability coefficients, dependability coefficients, standard errors of measurement, information functions, agreement indices—serve as evidence of broader social and scientific values that are critically important in assessment. So a reliability coefficient is a piece of evidence that operationalizes the values of accuracy, dependability, stability, consistency, or precision. In practice and in rhetoric, however, the methodologies for evidence reliability are often conflated with the social and scientific values of reliability. (p. 2)

If the methods cannot produce the evidence needed to support reliability, then we need to develop better methods. Parkes (2007) contended that reliability, like validity, needs to be considered as an argument. According to Parkes (2007), a reliability argument has six components, the first and most critical of these is determining the social and scientific values clearly. He argued that in constructing a reliability argument, assessment developers need to

1. Determine the social and scientific values (dependability, consistency, etc.) that are most relevant and decide which ones are most important.
2. Articulate clear statements of the purpose and context of the assessment, which includes making explicit the reasons the information is needed and how it will be used.
3. Define "replication" in the particular context, specifically structural versus conceptual replication.
4. Determine the "tolerance" or level of reliability needed.
5. Collect the evidence from the assessment, which may include traditional reliability data but it might also include other information such as narrative evidence.
6. Pull all of the information together to make the judgment and explaining how the evidence supports the final judgment. (pp. 6-7)

Parkes also emphasized that at the start, it is "easy to think of methods . . . rather than values" first but that it is "critical to stay focused on the value itself" and to determine which value or values are more important than others (p. 6).

Is consistency, for example, more important that stability? Or is precision more important than consistency? At this point, Parkes (2007) explained, it is very important to think about the construct being assessed, which introduces validity into the process. In other words, while reliability is distinct from validity, an appropriate argument for a context-specific form of reliability should be part of any validity argument.

While Parkes did not use Lakoff's concept of reframing, his approach helps us to reframe reliability in ways consistent with Lakoff because Parkes (2007) focused on values, which is what Lakoff recommended in reframing, and both called for articulating values and then using (or developing) methods that support those values. Parkes' (2007) approach to reliability also highlights the significance of purpose and context, which are critical components in effective communication and in assessing writing (Huot, 2002; CCCC, 2006). Haertel (2006) emphasized this point as well: "It bears repeating that in describing score accuracy, the statistics used and the ways they are interpreted must be suitable to the context and purpose of the measurement" (p. 67).

In supporting his approach to reliability, Parkes (2007) used an extended example of a classroom-based assessment of collaboration, performed by a classroom science teacher, to explain how reliability can—and should—work in performance-based assessments of complex, multi-dimensional activities. Using Parkes' (2007) position to reframe reliability in writing assessment would change the focus of the discussion from interrater reliability statistics to issues of purpose, context, evidence, tolerance, and effectiveness without dismissing reliability as unimportant, irrelevant, or impossible. Instead of asking what the statistics are for rater agreement, one might consider other questions, as Smith (1992, 1993) did. Smith reframed the question about reliability of the placement test results. Instead of looking exclusively at the interrater reliability statistic for the group, which was typical, Smith thought about agreement of raters in a much more nuanced way, examining raters' agreement with him/herself as well as within pairs of raters. He also looked at raters' disagreements to see if they were consistent. Ultimately, Smith's focus on reliability was considered in terms of the adequacy of placement: Were students adequately placed into the composition sequence? This reframing put the scoring reliability in the service of the validity of the placement exam results and situated it in terms of the particular writing program and course. Instead of "scoring essays," Smith had teachers placing students into the courses. He still wanted to be sure that students were being placed reliably—would the same student be put in the same course if the essay was read by another reader? By another pair of readers?-but he developed different methods for achieving reliable and valid results (O'Neill, 2003).

While Smith worked with single sample impromptu essays in developing his system, Haswell used single sample impromptus and portfolios to develop a two-tiered expert reader system for a Junior Writing Portfolio assessment program (Haswell & Wyche, 1996). The systems developed by Smith and Haswell, which were implemented over fifteen years before Parkes' (2007) essay, demonstrate one of Parkes' (2007) main points—that the focus of reliability needs to be on the values (such as accuracy, consistency and fairness) associated with reliability within the context of the assessment's purpose and context. By emphasizing this approach, new methods can be developed that produce both reliable and valid results. Parkes' (2007) framework for reliability can also help us communicate more clearly about a writing assessment so that it is framed by our values, purposes and theories.

CONCLUSION

Writing assessment scholars and practitioners have had significant influence in promoting performance-based assessments as well as in developing methods for scoring them (Lane & Stone, 2006). However, these assessment experts have not always been experts in language and literacy but in psychometrics and educational measurement. In many ways, writing specialists have been content to assign reliability and reliability methods to psychometricians, distancing ourselves from it. Parkes' (2007) contention, that reliability (like validity) needs to be considered as an argument, demands language and literacy experts to participate in discussions of reliability because constructing the reliability argument requires knowledge of more than psychometric statistics and methods. Reframing reliability to emphasize our values about writing, teaching writing, and learning to write will emphasize finding methods to build an effective reliability argument instead of merely reporting reliability co-efficients, which scholars have demonstrated to be problematic in writing assessment practice (Cherry & Meyer, 1993; Hayes & Hatch, 1999).

In Parkes' (2007) approach to reliability, writing assessment administrators would need to explain how reliability is being determined, why this approach is appropriate in the particular context, how specifically reliability is being calculated, the threshold for acceptable reliability and a justification for it, the limitations of the reliability, and how reliability contributes to the overall validation of the assessment's results. In determining reliability, many of us responsible for writing assessments should collaborate as equal partners with colleagues who have the statistical expertise. Writing assessment practitioners and scholars need to accept our responsibility to develop and maintain writing assessments that are informed by both language-based and psychometric theory and research.

We need to develop new methods for assessment as well as for determining reliability and validity if current methods do not work adequately for our purposes, as Parkes (2007) argued. This may mean collaborating with others who have different kinds of experiences and expertise, learning more about psychometric theory and practices, and engaging in difficult discussions with colleagues about what we value and why it matters.

By emphasizing values, we can begin to not only reframe reliability but also build more collaborative relationships with the educational measurement community. In writing assessment, this reframing can help writing teachers and administrators discuss and negotiate appropriate writing assessments with institutional administrators and others in more nuanced and effective ways. We must remember that validity and reliability connect to values such as accuracy, consistency, fairness, responsibility, and meaningfulness that we share with others, including psychometricians and measurement specialists. Focusing on these values and working to develop methods for upholding them can lead to the development of writing assessment methods that not only support teaching and learning but also are supported by evidence-based and theoretically-informed arguments. Over time, we will be able to shift the frame associated with reliability away from statistical methods and calculations to values that these methods—as well as methods not yet developed—should be supporting.

I believe we can be successful in our efforts to reframe reliability; after all, we were instrumental in resisting the move away from essay exams made in the 1940s, insisting that student writing needed to be evaluated in writing assessment. This position led to the development of holistic scoring and other methods for evaluating performance assessments (Huot & Neal, 2006; Lane & Stone, 2006). As scholars, teachers, and assessment practitioners, we need to engage in thoughtful ways to reframe reliability so that our assessments serve students and programs as they enact what we know about language and literacy.

REFERENCES

Allen, M. S. (1995). Valuing differences: Portnet's first year. *Assessing Writing*, *2*(1), 67-89.

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

Belanoff, P., & Dickson, M. (Eds.). (1991). *Portfolios: Process and product*. Boynton/Cook.

Black, L., Daiker, D., Sommers, J. & Stygall, G. (Eds.). (1994). *New directions in portfolio assessment: Reflective practice, critical theory and large-scale scoring*. Boynton/Cook.

Black, L., Helton, E., & Sommers, J. (1994). Connecting current research on authentic and performance assessment through portfolios. *Assessing Writing, 1(1)*, 247-266.

Borrowman, S. (1999). Trinity of portfolio placement: Validity, reliability, and curriculum reform. *Writing Program Administration*, *23(1-2)*, 7-27.

Broad, B. (2003). *What we really value: Beyond rubrics in teaching and assessing writing*. Utah State University Press.

Broad, R. L. (1994). "Portfolio scoring": A contradiction in terms. In L. Black, D. Daiker, J. & Stygall, G. (Eds.). *New directions in portfolio assessment: Reflective practice, critical theory and large-scale scoring* (pp. 263-277). Boynton/Cook.

Burke, K. (1966). *Language as symbolic action*. University of California Press.

Calfee, R. & Perfumo, P. (Eds.). (1996). *Writing portfolios in the classroom*. Lawrence Erlbaum.

Camp, R. (1993). Changing the model for the direct assessment of writing. In M. M. Williamson and B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 45-78). Hampton.

Cherry, R. D., & Meyer, P. R. (1993). Reliability issues in holistic assessment. In M.M. Williamson and B. A. Huot (Eds.) *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 109-141). Hampton.

Common Core State Standards Initiative. n.d. National Governors Association and Council of Chief State School Officers. http://www.corestandards.org/

Conference on College Composition and Communication. (Nov. 2006). Writing assessment: A position statement (Rev. ed.). National Council of Teachers of English. http://www.ncte.org/cccc/resources/positions/123784.htm

Daiker, D. A., Sommers, J., & Stygall, G. (1996). Pedagogical implications of a college placement portfolio. In E. M. White, W. D. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 257-270). Modern Language Association.

Diederich, P. B. (1974). *Measuring growth in English*. National Council of Teachers of English.

Durst, R. K., Roemer, M., & Schultz, L. (1994). Portfolio negotiations: Acts in speech. In L. Black, D. Daiker, J. Sommers & G. Stygall (Eds.) *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 286-300). Boynton/Cook.

Elbow, P. & Belanoff, P. (1986). Staffroom interchange: Portfolios as a substitute for proficiency examinations. *College Composition and Communication*, *37(3)*, 336-339.

Elliot, N. (2005). *On a scale: A social history of writing assessment in America*. Peter Lang.

Faigley, L, Cherry R. D., Jolliffe, D. A., & Skinner, A. M. (1985). *Assessing students' knowledge and processes of composing*. Ablex.

Gere, A. R. (1980). Written composition: Toward a theory of evaluation. *College English*, *42(1)*, 44-48, 53-58.

Godshalk, F. I., Swineford, F. & Coffman, W. E. (1966). *Measurement of writing ability*. (CEEB RM No. 6.) Educational Testing Service.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 65-109). ACE Praeger Series in Higher Education.

Hall, S. (1983). The narrative construction of reality. *Context*. https://www.academia.edu/34971009/Stuart_Hall_The_Narrative_Construction_of_Reality_1984_

Hamp-Lyons, L. & Condon, W. (2000). *Assessing the portfolio: Principles for practice, theory, and research*. Hampton Press.

Haswell, R. H. (Ed.). (2001). *Beyond outcomes: Assessment and instruction within a university writing program*. Ablex.

Haswell, R. H., & S. Wyche. (1996). A two-tiered rating procedure for placement essays. In T. W. Banta (Ed.), *Assessment in practice: Putting principles to work on college campuses* (pp. 204-207). Jossey-Bass.

Haswell, R. H., Johnson-Shull, L. & Wyche-Smith, S. (1994). Shooting Niagra: Making portfolio assessment serve instruction at a state university. *Writing Program Administration*, *18,* 44-54.

Hayes, J. R. & Hatch, J. (1999). Issues in measuring reliability. *Written Communication*, *16*, 354-367.

Hertog, J., & McLeod, D. 2001. A multiperspectival approach to framing analysis: A field guide. In S. D. Reese, O. H. Gandy, & A. E. Grant (Eds.), *Framing public life* (pp. 131-161). Lawrence Erlbaum.

Hester, V., O'Neill, P., Neal, M., Edgington, A., & Huot, B. (2007). Adding portfolios to the placement process. In P. O'Neill (Ed.), *Blurring boundaries: Developing writers, researchers, and teachers* (pp. 61-90). Hampton Press.

Hoetker, J. (1982). Essay examination topics and student writing. *College Composition and Communication*, *33(4),* 377-392.

Huot, B. (2002). *(Re)Articulating writing assessment for teaching and learning*. Utah State University Press.

Huot, B. & Neal, M. (2006). Writing assessment: A techno-history. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 417-432). Guilford Press.

Kearns, E. (1993). On the running boards of the portfolio bandwagon. *Writing Program Administration*, *16(3)*, 50-59.

Kuhn, T. S. (1962). *Structure of scientific revolutions*. University of Chicago Press.

Lakoff G. (2002). *Moral politics: How liberals and conservatives think* (2nd ed). University of Chicago Press.

Lakoff, G. (2004). *Don't think of an elephant! Know your values and frame the debate*. Chelsea Publishing.

Lakoff, G. (2006) Simple framing. *Rockridge Institute*. http://www.rockridgeinstitute.org/projects/strategic/simple_framing

Lane, S. & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.) *Educational measurement* (4th ed). (pp. 387-431). ACE Praeger Series in Higher Education.

Larson, R. L. (1996). Portfolios in the assessment of writing: A political perspective. In E. M. White, W. D. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practice*. (pp. 271-283). Modern Language Association.

LeMahieu, P. G., Eresh, J. T., & Wallace, R. C. (1992). Using student portfolios for a public accounting. *School Administrator*, *49*(11), 8-13.

LeMahieu, P. G., Gitomer, D., & Eresh, J. (1995). Portfolios in large scale assessment: Difficult but not impossible. *Educational Measurement: Issues and Practice*, *14(3*), 11-28.

Lowe, T. J., & Huot, B. (1997). Using KIRIS writing portfolios to place students in first-year composition at the University of Louisville. *Kentucky English Bulletin*, *46*, 46-64.

Lynne, P. (2004). *Coming to Terms: A theory of writing assessment*. Utah State University Press.

Messick, S. (1989). Meaning and value in test validation: The science and ethics of assessment. *Educational Researcher*, *18(2)*, 5-11.

Mislevy, R. J (2004). Can there be validity without "reliability?" *Journal of Educational and Behavioral Statistics*, *29(2)*, 241-245.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*, *23*(4), 5-12.

Murphy, S. & Underwood, T. (2000). *Portfolio practices: Lessons from schools, districts and states*. Christopher Gordon.

Nelson, A. (1999). Views from the underside: Proficiency portfolios in first-year composition. *Teaching English in the Two Year College*, *26*, 243-253.

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

O'Neill, P. (2003). Moving beyond holistic scoring through validity inquiry. *Journal of Writing Assessment*, *1(1)*, 47-65. https://escholarship.org/uc/item/4qp611b4

Parkes, J. (2007). Reliability as argument. *Educational Measurement: Issues and Practice*, *26*(4), 2-10.

Peckham, I. (2009). Online placement in first-year writing. *College Composition and Communication*, *60(3)*, 517-540.

Penrod, Diane. (2005). *Composition in convergence: The Impact of new media on writing assessment*. Lawrence Erlbaum.

Roemer, M., Schultz, L. M., & Durst, R. K. (1991). Portfolios and the process of change. *College Composition and Communication*, *42(4)*, 445-469.

Shale, D. (1996). Essay reliability: Form and meaning. In E. M. White, W. D. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices*. (pp. 76-96). Modern Language Association.

Smith, W. L. (1992). The importance of teacher knowledge in college composition placement testing. In J. R. Hayes (Ed.), *Reading empirical research studies: The rhetoric of research*. (pp. 289-316). Ablex.

Smith, W. L. (1993). Assessing the reliability and adequacy of using holistic scoring of essays as a college composition placement program technique. In M. M. Williamson and B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 142-205). Hampton Press.

Sommers, J., Black, L., Daiker, D., & Stygall, G. (1993). Challenges of rating portfolios: What WPAs can expect. *Writing Program Administration*, *17(1-2)*, 7-29.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9*.

U.S. Department of Education. Race to the Top Fund. (n.d). http://www2.ed.gov/programs/racetothetop/index.html

Underwood, T., & Murphy, S. (1998). Interrater reliability in a California middle school English/Language Arts portfolio assessment program. *Assessing Writing*, *5*(4), 201-230.

White, E. M. (1993). Holistic scoring: Past triumphs and future challenges. In M. M. Williamson and B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 79-108). Hampton Press.

White, E. M. (1994). *Teaching and assessing writing* (2nd ed). Jossey Bass.

White, E. M. (1995). Apologia for the timed impromptu essay test. *College Composition and Communication*, *46(1)*, 129-139.

White, E. M. (2005). The scoring of writing portfolios: Phase 2. *College Composition and Communication, 56(4),* 581-600.

Whithaus, C. (2005). *Teaching and evaluating writing in the age of computers and high-stakes testing*. Erlbaum.

Wiggins, G. (1993). Constant danger of sacrificing validity to reliability: Making writing assessment serve writers. *Assessing Writing*, *1(1)*, 129-139.

Willard-Traub, M., Decker, E., Reed, R., & Johnston, J. (1999). Development of large-scale portfolio placement at the University of Michigan 1992-1998. *Assessing Writing*, *6*(1), 41-84.

Williamson, M. M. (1993). Introduction to holistic scoring: The social, historical, and theoretical context for writing assessment. In M. M. Williamson and B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 1-43). Hampton Press.

Williamson, M. M. (1994). Worship of efficiency: Untangling theoretical and practical considerations in writing assessment. *Assessing Writing*, *1,* 147-174.

Williamson, M. M. (2003). Validity of automated scoring: prologue for a continuing discussion of machine scoring student writing. *Journal of Writing Assessment*, *1(1)*, 85-104. https://escholarship.org/uc/item/8nv3w3w8

Yancey, K. B. (Ed.) (1992). *Portfolios in the writing classroom: An introduction*. National Council of Teachers of English.

Yancey, K. B. (1999). Looking back as we look forward: Historicizing writing assessment. *College Composition and Communication*, *50(3)*, 483-503.

Yancey, K. B., & Weiser, I. (1997). *Situating portfolios: Four perspectives*. Utah State University Press.