CHAPTER 8.

VALIDITY OF AUTOMATED SCORING: PROLOGUE FOR A CONTINUING DISCUSSION OF MACHINE SCORING STUDENT WRITING

Michael Williamson
Indiana University of Pennsylvania

Writing assessment has developed along two separate lines, one centered in professional organizations for writing teachers and the other centered in professional organizations for the broader assessment community. As the controversy about automated scoring continues to develop, it is important for writing teachers and researchers to become fluent in the discourse of the broader assessment community. Continuing to label the work of the broader assessment community as positivist and continuing to ignore it will only result in a continuing sense of defeat as automated assessment is adopted more widely. On the other hand, an examination of the literature on educational assessment will reveal that the theoretical base for assessment is quite consistent with the principles adopted by the writing assessment community.

> Grading essays by computer seems to have entered an explosive new phase, and I hope that, by the end of this talk, you folks will be excited, too, about all the changes this may mean for testing. After all, essay grading has been done for perhaps 4 thousand years. But now we seem to face a brand-new opportunity: Not simply to help in human essay grading, but to firm it up with actual objective data, of the kind never really used.
>
> – Ellis Batten Page (1995)

Anson (2003), reflecting on developments in artificial intelligence (AI), suggests it has provided little to serve any useful purpose in the English classroom because software has not been sufficiently sophisticated. Earlier, Herrington and

DOI: https://doi.org/10.37514/PER-B.2024.2166.2.08

Moran (2001) examined an emerging application of AI in English Studies, automated scoring, and the use of computer algorithms to simulate holistic ratings of student writing. Although they are concerned about the adequacy of the feedback provided by such programs, the greater concern is the implications for students' learning when computers are the basis for grades. However, automated scoring technologies are finding wider acceptance among educators. The Commonwealth of Pennsylvania recently made a commitment to the use of Intellimetric, the scoring engine reviewed by Herrington and Moran (2001). Some reports suggest that this engine was to be used in 2003 to score the writing of students on the mandatory Pennsylvania state achievement examinations (Indiana Gazette, 2003). Other states and individual school districts are either implementing or exploring the implementation of one of the available engines.

This obvious conflict suggests that some may see valid applications for automated scoring, whereas others see none, suggesting that a deeper examination of the available inquiry about the validity of automated scoring is necessary. English teacher response to automated scoring has been limited and such response (Anson, 2003; Herrington & Moran, 2001) does not refer to any of the evidence presented by the developers of automated scoring programs. There remains a need to examine the claims made by test developers about the validity of automated scoring and to determine whether any possible objections have been addressed.

Initially, I hoped to write an article that picked between the various arguments and claims and contended for certain use of automated scoring in writing assessment. Unfortunately, my reading of the literature around this issue left me feeling that other precursor work needed to be done before the two camps, what Moss (1998) first labeled college writing assessment and educational measurement, could productively learn to talk to each other about automated scoring. In this article, I explore various beliefs and assumptions held by each side. Looking at the history of test development in general and writing assessment in particular, I examine the drive toward more reliable and efficient ways to measure educational achievement and writing ability. Additionally, I consider the various epistemological orientations of those who work in social science and the humanities, noting how each disciplinary area has changed over the last several years with the influence of postmodern theories of knowing and making meaning. I hope that this article can establish a common ground for future scholarship and discussion. At the very least, automated scoring is an incredible research opportunity through which we can explore the many different ways student writing can be read, valued, and sanctioned.

Automated scoring is not new. It first appeared in 1966, in the work of Ellis Page (1995). The response to this early work from the English-teaching community was similar to current responses. Reviewed in *Research in the Teaching of*

English, Page's original work drew a response similar to Anson and Herrington and Moran from Macrorie (1969). On the other hand, Coombs (1969) was skeptical, but not entirely dismissive of the potential demonstrated in Project Essay Grade. However, automated scoring does not seem to have been wholeheartedly embraced by anyone in English Studies publishing in typical outlets, such as *College English* or *Research in the Teaching of English*.

On the other hand, a recently burgeoning literature on automated scoring has appeared in the literature typically examined by the broader assessment community, much of it suggesting that automated scoring does have valid applications for the assessing of writing.

AUTOMATED SCORING AS WRITING ASSESSMENT

Although it has a new face, the controversy over automated scoring reflects the constant struggle over writing assessment and the apparent stasis in achieving a resolution (Williamson, 1993). Until recently, the controversy focused on movement from indirect to direct measurement of writing (Williamson, 1993), as reflected in Yancey's (1999) history of the last 50 years of writing assessment. Currently, writing assessment seems to be caught in a three-way tug of war involving the introduction of portfolio assessment in the teaching and assessing of writing. Yancey suggests a shift in focus from reliability in the dispute over direct and indirect assessment, to validity, a dispute over how much writing is necessary to make a valid judgment about students' writing. From the beginning, there has been an explicit concern about the effects of particular approaches to assessment on the teaching and learning of writing, in effect, a question about the validity of assessment. Yancey's view reflects a trend in the literature by and for writing teachers and researchers to respond primarily to the challenges posed by systems developed to ensure the reliable scoring of student writing. The proposal to replace essay examinations with objective examinations, based in multiple-choice technologies began the controversy.

> The most recurrent criticism of essay tests, and the one about which the most has been written, concerns the unreliability of evaluating essay answers. If a test is to be worth while [sic] as a measuring instrument, it must measure what it purports to measure consistently and dependably (Stalnaker, 1951, p. 498).

As Yancey points out, the response to objective testing was the development of direct assessment approaches using writing, justified in terms of their reliability, just as the justification for indirect assessment, using multiple-choice items, was grounded in its reliability compared to the earlier use of writing as a tool for assessment. Although the battleground itself was seen as reliability, the larger struggle was about validity, thought it was focused at the time in terms of reliability.

All educational measurements are generally intended to elicit information regarding the structure, dynamics, and functioning of the student's mental life as it has been modified by a particular set of learning experiences. The special problem in the case of the achievement test is to obtain information which is reliable and pertinent, and to do so efficiently (Stalnaker, 1951, p. 496). These concerns evolved into the traditional claim that a test had to measure what it purports to measure and that reliability is a necessary but insufficient claim for validity.

Validity has two aspects, which may be termed relevance and reliability. "Relevance" concerns the closeness of agreement between what the test measures and the function that it is used to measure. "Reliability" concerns the accuracy and consistency with which it measures whatever it does measure in the group with which it is used. To be valid—that is, to serve its purpose adequately—a test must measure something with reasonably high reliability, and that something must be fairly closely related to the function it is used to perform (Cureton, 1951, p. 622).

Although some developers have made claims about potential pedagogical uses of automated scoring programs, I only focus on their validity as it pertains to writing assessment. The larger, and perhaps more important issue of their pedagogical value is another question, one that does not seem of immediate relevance for writing assessment. I begin with Herrington and Moran's (2001) exploration because it reflects my own examination of particular programs. There is, however, a paucity of research beyond such informal examinations. Second, for the most part, feedback to students is based on boilerplate rubrics, some quite complex and sophisticated. Rubric-based feedback in any kind of scoring may not address the particular reason an essay was placed in a score category (Broad, 2003; Huot, 1993; Pula & Huot, 1993; Smith, 1993). The qualities and bases for human judgment of complex performances cannot be explained by a rubric. 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.

VALIDITY

As early as 1951, validity was defined by Edward Cureton in the first edition of what would become a periodic definition of the state of the art in educational measurement, *Educational Measurement*.

The essential question of test validity is how well a test does the job it is employed to do. The same test may be used for several different purposes, and its validity may be high for one, moderate for another, and low for a third (p. 621).

An important and forward-looking aspect of this definition is that it is grounded in the use of a test, not in the test itself. The definition of validity evolved with both formal and informal meanings, as can be noted in Cronbach's (1971) leading text on the theory and practice of educational measurement.

We defined validity as the extent to which any measuring instrument measures what it is intended to measure. However, as we pointed out in Chapter 1, strictly speaking, "One validates, not a test, but an interpretation of data arising from a specified procedure" (p. 447).

While conforming with this general definition, Anastasi (1976) presents three primary forms of validity, each defined by the procedures used to determine them.

Fundamentally, all procedures for determining test validity are concerned with the relationships between performance on the test and other independently observable facts about the behavior characteristics of under consideration (p. 134).

She also provides a separate treatment of validity as an issue for interpreting test results through the use of decision theory, a further coupling of validity with particular uses of a test.

In one of the seminal works on writing assessment produced by writing researchers, Cooper and Odell (1977) define validity with a slightly different focus, one that may ultimately be responsible for spreading the informal definition of validity as the dominant meaning in writing assessment.

If a measure or measurement scheme is valid, it is doing what we say it is doing. We want to insist on a careful distinction between predictive validity and other kinds of validity, content and construct validity (p. xi).

This definition reflects what I am labeling the informal definition of validity. Later definitions of validity tend to adopt this informal definition, for instance:

> Although validity is a complex concept—colleges offer advanced courses in it—one simple concept lies behind the complexity: honesty. Validity in measurement means that you are measuring what you say you are measuring, not something else, and that you have really thought through the importance of your measurement in considerable detail. (White, 1994, p. 10)

White's definition of validity is metaphorical, and although metaphor is not unknown in social sciences research, the redefining of validity in this case moves two fundamentally different definitions of the same concept further apart.

The essential, crucial difference between these two definitions lies in the distinction between defining validity as procedure and validity as a property of a test. This distinction emerges from the difference between understanding the mathematical basis for assessment and the application of assessment in what

Stalnaker (1951) labels achievement testing. Tests like statistical operations are conducted to make informed educational judgments. The simplicity of distinction between a procedural and conceptual understanding of validity is not always as clear and separate as it might seem. The fundamental nature of validity can be rendered confusing by educational researchers themselves.

While the definition of validity seems simple and straightforward, there are several different types of validity that are relevant in the social sciences. Each of these types of validity takes a somewhat different approach in assessing the extent to which a measure measures what it purports to (Carmines & Zeller, 1980, p. 17).

Broad (2003) labels one stance in writing assessment "positivist," a stance that can be traced to Berlin's (1984) history of writing instruction. Positivism as a theoretical approach to the philosophy of science certainly characterizes early psychometric theory and its attempt to define psychology and educational and psychological assessment as a science. Guilford (1954) traces the emergence of statistical investigation in psychology and grounds his approach to the field in mathematics, as well as statistical inquiry, "The progress and maturity of a science are often judged by the extent to which it has succeeded in the use of mathematics" (p. 1). Gulliksen (1950) specifically limits his description of mental testing to those defined by quantitative methods, while specifically noting the difference between statistics and mathematics. In Guilford's terms, mathematics is a "universal language that any discipline may use with power and convenience" (p. 1). That this movement toward the use of mathematics and quantification may be positivist is one that deserves larger exploration in the literature of the field. However, there is an interesting contrast to what may be perceived as the problem of quantification in writing assessment.

As early as the 1950s, at least, such issues as validity were seen less as defined by the results of a statistical test than as a matter of disciplinary disputation, the assembling of evidence, not the simple results of a statistical test (Cureton, 1951). In a related example, in discussing educational evaluation, one of the primary applications of educational measurement, Cooley and Lohnes (1976), both eventually to become president of the American Educational Research Association, suggest that the scrutiny of the field and not objectivity is the issue. Moss (1998) calls her response essay to a study of writing assessment validation, "The Test of the Test." For Moss, validation is a practice in turning the gaze toward the construct of the assessment itself. It is a form of reflective practice, or as Ellen Schendel (1999) claims, "social action."

What tends to keep researchers honest is the publicly available record of what they did and what they found, and not a godlike objectivity which some people seem to feel those doing evaluations should exhibit. Scientists doing basic

research know that if their work is to have any value whatsoever, it will be closely read and critically examined by their colleagues in the field (Cooley & Lohnes, 1976, p. 2).

These and other perspectives of validity are rooted in the ideas of Cronbach (1988, 1989) and Messick (1989). Cronbach (1989) characterizes validity as a form of disciplinary argumentation, one that is never finished and that evolves with each new use of an assessment in a new locale: "Validation is a lengthy, even endless process" (p. 151). Such a definition is supported by Cureton (1951) and Anastasi (1976) as well. It is this definition that leads Huot (2002) to characterize assessment as a continuing form of research. Thus, writing assessment should be viewed as a continuing examination of the available tools for assessment, as they are used for making new decisions. New developments will inevitably bring new tools, all of them requiring validity inquiry of their own.

Smith (1993) is probably the first researcher in writing assessment who fully reflects the complexity of validity inquiry. Although his work is some of the first substantive research that looks at the validity and not reliability of a writing assessment (Huot, 1996), ironically, he eschewed the word validity because he wanted to avoid any baggage associated with such a term. He used accuracy of placement as the goal for his placement testing program at the University of Pittsburgh. With collaborators, he designed a series of studies on the procedures that structure the way teachers make decisions based upon their reading of placement essays. Each of the studies led to a modification of the procedures that allowed a stronger claim to the validity of the assessment, the accuracy of placement of students in the writing program. This not only demonstrates more accurate placement of students over time, but it also led to a modification of the scoring procedures themselves. The end result was a less costly system because the reading and decision making were rooted in the context about which the teachers were expert.

The notion of validity as argument and the nature of professional judgment is related to Bleich's (1975) view of interpretive communities and Kuhn's (1996) view of the way that science changes through changes in the worldview of the members of the discipline. The meaning of a text, be it a poem or a validity inquiry, lies with the community of readers in the field and their intertextual experiences with the field. Such a position reflects a more postmodern view than the positivism cited as the basis for psychometric theory.

An additional consideration for validity is the impact of the assessment (Messick, 1989). The consequences of decisions made on behalf of a test is a core concern for validity inquiry because the use of a test may impact what is learned and how that learning takes place. This concern for the impact of a test is one of the ethical bases for validity theory. Thus, validity inquiry must examine how

learning changes as a result of the implementation of an assessment. Although this sounds like an ethical way to proceed, English professionals might question the existence of studies of the consequences of high-stakes testing on individual students taking high-stakes tests. Interaction among various fields is important if we are to understand complex phenomena. In particular, measurement theory in education and psychology has to respond to developments in psychology and education if the field is to remain viable. The impact of theoretical changes is not universally distributed in a field (Kuhn, 1996). If there are specialists in educational measurement still working with a variety of validities, there are still writing teachers and researchers who pursue grammar study as a prescriptive methodology in the teaching of writing. If validity theory has not coalesced into a univocal stance in measurement, the meaning of error is equally problematic for many teachers who are not able to grasp or who are unfamiliar with the complexity of disciplinary discourse on error.

After all, members of any academic field are part of both the paradigm that is disappearing and the new paradigm that provides a new synthesis for the field (Kuhn, 1996). That some may quote the contemporary definition and unwittingly include older definitions is not surprising. An interpretive community does not need to be, indeed is unlikely to be, univocal about any reading of any text. Importantly, if early theories of assessment were deterministic in the positivist sense that they were seen as objective explanations of reality, the postmodern influence in assessment publicly acknowledges the debate that always existed, and provides a new understanding of the meaning of such debate.

The core of my concern in the different representations of validity has to do with the difference between English Studies and educational measurement, the difference between social science and humanistic disciplines. A science depends on a clearly defined methodology as the basis for disciplinary disputation. Although English Studies depends largely on a hermeneutic form of inquiry, one based in close reading, assessment depends on evidence defined by the procedures that are used to collect it. For instance, the heart of the definition provided by Carmines and Zeller (1980) highlights the defining of each of the various types of validity as a procedure, despite the fact that it misses the more important concern that validity is contextualized.

Two conflicting views of research methodology are the primary problem for humanists as they attempt to represent their views outside of English Studies because any argument about validity will have to face the need to address the basic procedural issues of social science. Furthermore, if validity is seen as a unitary construct that involves the consequences of the test's use in context, validity can be seen as a situated construct, one that must observe the same situatedness that literacy theorists have been articulating for some time.

As a student of English Studies, I am concerned by the claims of Herrington and Moran (2001). As any good scholars in the field, they read the text of the automated scoring engine and see the rhetorical implications of its use in English classrooms. However, as a student of educational assessment, I know that their review of automated assessment does not provide the kind of structured inquiry necessary to convince a member of the community of readers in assessment. It is easy to adopt the stance that all psychometricians are positivists if one does not understand the fundamental role of scientific procedure in defining inquiry. However, the label itself has no meaning outside of English Studies because any form of quantification is labeled positivist. The label itself is, therefore, one that does not make the case against claims by psychometricians about the validity of particular approaches to assessment. In fact, most first-year composition texts would probably characterize such an argument as ad hominem.

PRINCIPLES GUIDING THE EXAMINATION OF THE VALIDITY OF AUTOMATED SCORING

All of the following statements are derived from the literature on education assessment and follow from my characterization of validity as it is defined by the following:

1. The validity of an assessment lies in the decision that is made on the basis of the test, not the test itself.
2. Validity is a form of scholarly argumentation, based on research, which subjects the assessment to open discussion about both its substance and its meaning.
3. Validity is not a substantial or concrete set of claims, the argument is open to question with each use of the assessment and as developments in various theories, both within and outside of assessment provide new perspectives on assessment, what is being assessed, and how the assessment is being used.
4. Validation research is never a closed circle. Each use of an assessment, whether in the same or different contexts must be examined to ascertain and revalidate the validity argument for the assessment, its uses, and the meaning of its uses.
5. In addition to examining the adequacy of the assessment for the decisions that are to be made from its use, assessment developers and users also have an ethical responsibility to examine the consequences of an assessment, to examine the effects of the assessment on both immediate contexts and broader cultural contexts.

Notice that each of these statements contains a procedural definition of validity. I argue that the definitions of validity that are common in English Studies are static, indeed, are positivist in the sense that they suggest we can know that a test is valid in objective terms, because we can know it is doing what we say it is doing. In other words, because many in English Studies ascribe to an older notion of validity (White, 1994; Yancey, 1999), they are unwittingly missing an opportunity to apply postmodern theories to validity inquiry and are, instead, promoting a rigid, decontextualized "positivist" concept of validity for writing assessment.

ARTIFICIAL INTELLIGENCE

Automated scoring is based in the technology of AI, and claims to bring the relative efficiency of automation to scoring essays. These two concepts need to be defined as part of the process of validity inquiry. AI is a research paradigm built around several sciences. The primary goal of the emergent paradigm has been the simulation of human intelligence and behavior in the electronic system of a computer. Developments in each of these sciences, from linguistics to psychology and mathematics to computer science, have allowed a nearly continuous development of demonstrations of intelligent machines. The emergent technologies have resulted in a variety of applications that both enhance and simulate human performance in a variety of fields. Thus, it seems that the use of such technologies would inevitably lead to their application in English studies. The first such application—Project Essay Grade—was seen by its developers as a method of relieving writing teachers of the burden of grading, leading also to more objective grades (Ajay, Tillett, & Page, 1973; Page, 1966, 1967a, 1967b, 1995; Page & Fisher, 1968). After an initial ambiguous response (Coombs, 1969; Macrorie; 1969), the concept of computer grading seems to have had little attention from researchers in composition and rhetoric for some time (Huot, 1996).

The development of the personal computer in the 1980s led to an outburst of enthusiasm for the use of computers in the writing classroom. The cutting edge of the field of computers and composition was initially defined by the seminal work of Hugh Burns (1979) with rhetorical invention and the rapid growth of word processing, among other business and personal applications. Burns' work reflected the early applications of artificial intelligence to English Studies. His work demonstrated the programming theories of artificial intelligence pioneered by Joseph Weizenbaum in the development of Eliza, a computer program designed to simulate the psychotherapeutic interviews of Carl Rogers. Eliza was considered to be a failure because the program did not meet Turing's criterion for a computing machine simulating human behavior,

a primary consideration in judging the validity of computer programs that "artificially" simulate human intelligence.

Alan Turing was one of the pioneers of digital computing at Bletchley Park in England during World War II. As a very early theorist in computing, he suggested that a successful demonstration of human intelligence by a computer would be indistinguishable from the performance of an actual human. In other words, Liza would be successful if the program were able to provide counseling to a human client without the client being able to determine whether the advice came from a machine or another human. Neither *Eliza* nor Burn's invention programs meet the criterion because they were unable to respond coherently to aberrant statements. The result of aberrant statements or questions about questions from the human user resulted in meaningless responses from the programs. Although the programming had a rudimentary syntactic parser, enabling it to extract relevant words from the input, it had no means of examining the meaning of any of the input. Therefore, it was easily "fooled" into giving unintelligible or meaningless responses. Subsequently, demonstrations of AI have been based on successively sophisticated approximations of human intelligence. Most of these early demonstrations were intended only to model what was possible, not necessarily to meet Turing's criterion.

Since the early demonstrations of machine intelligence, researchers working in the multidisciplinary field of natural language processing were busy with both basic research into computer simulation of language and immediate applications of this technology. With each new demonstration of the emerging technology, more sophisticated responses to human language were possible, as were more sophisticated applications. The accessibility of computers to those outside of computer science owes as much to the developments in AI as to the developments in the electronics side of computing.

AUTOMATION

Automated scoring—the use of computers to simulate holistic ratings of English essays—is quite accurately described as automation in the original sense of the word—the use of technology to relieve humans of repetitive work, work that taxes the limits of our abilities. It is, simply, the performance of tasks by machines, tasks that were originally performed by skilled humans, made skilled humans more productive, or created less skilled work from more complex work. Early automation is represented by the agricultural machines that first improved tilling the soil and subsequently harvesting. The original Luddites of 1811-1812 were weavers in England, members of a craft guild who attempted to destroy the newly invented machinery that left fewer jobs for unskilled workers. Mechanical

developments in automation began to skyrocket with the introduction of computer technology. Today, labor unions representing the interests of workers have been watching the emergence of automation with considerable concern because industrial, production line workers have been replaced by electronically operated machines that perform repetitive tasks with greater precision and accuracy than humans, at least in the view of industries that have adopted this technology. The motivation underlying electronic automation, even as it was in the planning, viewed constant repetition as a weakness in humans. Industrial automation was motivated by efficiency. To the extent that computers can make any task more efficient, they will be of interest in industry. Although workers in AI may not perceive the impact of their work, much of the research and development for applications of the emergent theories in AI have been funded by governments and industry looking for ways to operate more efficiently, even if only to get beyond errors and other problems that reflect the limits of human performance.

In the case of automation, the concern for a computer's performance is not on whether it meets Turing's criterion. Instead, the question is whether the task itself is computable. According to Johnson-Laird (1977), computability depends on being able to specify a task with sufficient precision to develop a programming algorithm, based in the computational structure of computer software. For instance, welding an exact spot on a car body involves only a question of space and time—the movement of the machine to the location of the weld and the length of the welding time. Although the relative quality of human labor and automation is certainly one issue, the real question lies with the sufficiency of the performance of the machine. If sufficient quality can be achieved by a computer program or robot, operating at greater speed and less cost, clearly, the programming is successful. The cost reduction and increased efficiency of machine operation, when seen only in terms of the costs of production and profit margins, are clearly a business issue.

CAN HOLISTIC SCORING BE AUTOMATED?

In an earlier essay, I discussed in some detail the underpinning of much assessment practice in the "Worship of Efficiency" (Williamson, 1994). Further demonstration of the role of efficiency in assessment is provided by some of the sources cited earlier in this text (Cureton, 1951; Stalnaker, 1951). The question of validity for automated scoring turns, in this circumstance, on whether automated scoring can provide results at least as trustworthy as human raters with greater efficiency and less cost. From this perspective, the question of validity for automated scoring can be answered in the same way that questions of quality are determined for other forms of automation. Although cost accounting may

be more relevant to business, the mathematical apparatus of assessment theory is employed in demonstrating the quality and validity of automated assessment as it compares to holistic raters. For others, the question of any automation of the work of writing teachers and assessors is a question of the computability of human language in the first place. In other words, can a computer using AI *read*?

Validity, as it is related to the comparability of holistic rating, is considerably more limited than some of the larger questions that have been raised about holistic scoring itself, such as the adequacy of the criterion definition of writing represented by a single essay and the adequacy of the criterion definition of reading represented in standardized rating sessions. The distinction between these two views of validity inquiry about automated scoring has important consequences for how specific investigations into its validity will be understood.

For example, the key issue for those creating automated scoring is whether the program can predict holistic ratings of more than six raters (Burstein, 2003), many more than the number typically employed in a holistic scoring session. To support their claim, automated scoring needs to demonstrate that it is more efficient and costs considerably less than rating sessions.

However, the discussion, within English Studies, seems to be dominated by a very different definition of the activity of holistic rating. The criterion that Herrington and Moran (2001), as well as Anson (2003) appear to be using is whether a computer can read. At least three studies (Huot, 1993; Huot & Pula, 1993; Wolfe, 1997) established that holistic scoring is a limited form of reading. In the Huot and Pula and Huot studies, holistic raters made rapid decisions about the placement of students reflected in the writing, and then spent time responding to other aspects of the writing. Wolfe found that raters who agree at a high rate with each other have a more focused reading process.

For a social scientist, the immediate question is whether the procedures used by automated scoring engines simulate the scoring process of human raters. This question is more difficult to answer because holistic rating is not reading as is usually defined in literacy research where the goal is to produce various readings; the push for writing assessment has been toward a single reading (Elbow & Yancey, 1994). Holistic scoring, by definition, limits the features of a text that the rater attends to. The scoring process also limits the purposes for which a text is read. Such convergent reading is not what is typically represented as fluent adult reading, an act of making meaning that typically leads to divergent views of a text.

IS HOLISTIC SCORING VALID?

There have been two large studies of the validity of holistic scoring, as applied to individual essays (Gottshalk, Swineford, & Coffman, 1966) and to multiple

essays from the same writers, intended as a form of portfolio assessment (Breland et al., 1987). The earlier study suggested that multiple-choice tests of grammar predict a student writer's performance more accurately than an essay when it is scored using a holistic procedure by two or three raters. Consequently, the relatively cheaper and more efficient indirect approach was justified because it could predict an individual's score on a criterion with the greater precision and accuracy than a writing sample. The claim for the validity of indirect assessment is based on a form of criterion validity known as concurrent validity that compares an examinee's performance on two different valued measures. Veal and Hudson (1983) dispute that result in another study of the use of holistic scoring using state assessment data from Georgia in which students' performances on multiple-choice tests of usage and grammar do match well with a holistic score.

Breland et al. (1987), stipulating that direct writing assessment is more efficient and less costly, demonstrated that one essay read two or three times could attain the reliability of indirect measures. Their criterion definition of writing was six essays from each writer. They conclude that the best approach to writing assessment is a combination of both direct and indirect assessment because the two work together to provide both a broader and more reliable picture.

Although psychometric theory clearly supports the need for studying validity in particular applications of a test, in practice multiple-choice tests were considered adequate when used "off the shelf" by educational institutions. Thus, although the theory was suggesting the need for more study of assessment procedures in particular applications, conventional wisdom allowed for their use as ready made instruments for student, teacher, and program evaluation. This was equally true in the use of holistic scoring. For the most part, writing assessments used holistic scoring without much examination of the validity of its actual use, because the understanding of assessment theory prevalent in the field was that a test using writing is more valid on its face and in its content than any form of indirect test (Yancey, 1999).

White (1994) recounted the political struggles involved with the adoption of direct assessment. However, the extant theory in measurement could have been used to support the argument against indirect assessments had more writing assessment developers, like Veal and Hudson (1983), used the theory to argue their position. Hence, with greater fluency in the theory that was used, writing teachers and researchers would likely have been able to develop assessments that could be demonstrated to have the same kinds of properties that were valued in the validation of indirect assessment. One good example is the study by Breland et al. (1987), which suggests that holistic scoring of a writing portfolio leads to more accurate predictions than the score of any single essay in the portfolio. As early as Terman's 1916 book on the measurement of intelligence, statistical

procedures were well defined for an examination of the contributions of test length to overall test reliability. Item validity was really the only focal concern, because as Gulliksen (1950) points out:

> We see that the validity coefficient is the square root of that for the reliability coefficient. . . . Since the validity coefficient is usually considerably smaller than the test reliability, this usually means that changing the length of the test can be expected to have a very slight effect on the validity of the test. (p. 90)

Thus, the ultimate focus in measurement theory is on reliability to the extent that it is defined statistically. A test reliability of 0.9 will provide a test validity of 0.3, for instance. Little wonder that the traditional debate over holistic scoring confuses reliability and validity.

If, as I have argued, validity is seen as existing in a particular use of a test, in a particular context, at a particular time, validity reflects the situatedness of literacy as most researchers and teachers of writing have been claiming. Thus, validity does not lie in statistical procedures alone. However, test developers themselves rarely study the validation of decisions. Furthermore, the kind of study undertaken by Smith (1993) is costly and lengthy, and requires both experience and training in empirical research. Because efficiency is valued in applications of assessment theory (Williamson, 1994) and not very many involved with writing assessment have training in empirical research (White 1994), it is not surprising that there is very little validation research available for particular uses of writing assessment. Exceptions are seen in the work of Blakesley (2003) with Directed Self-Assessment and Herrington on the use of technology using Smith's (1993) and Haswell's (2001) approach to scoring.

VALIDATION STUDIES OF AUTOMATED SCORING

There is really only a single automated scoring engine that has a consistent record of validation research, eRater as it is used to score essays for the Graduate Management Aptitude test. Until recently, the essay portion of the test was read by a group of holistic raters, trained by Educational Testing Service (ETS), the test developer and vendor. The scores are used by graduate programs in business to determine admission to their programs. Like the SAT and the Graduate Record Examination (GRE), the scores are used as one indication of performance in a program of study, along with other indicators, such as class rank, grade-point average (GPA), and the school graduating the applicant. However, the responsibility of the actual validation of each of those examinations lies with the institutions that use them to make decisions about admissions. ETS cannot provide

validation data for any of those examinations because they do not have relevant local data to determine the suitability of each examination for the decision to admit or deny admission to an applicant to a particular program. Validation data, such as national norms and performance of students with self-reported characteristics such as GPA are frequently part of these examinations. But, the only place to determine the validity of admissions decisions is within the institution using the scores. In the case of the SAT, most admissions departments use the scores in formulas to predict such things as first-year GPA. Similarly, ETS reports the success of similar predictions for a number of schools as part of their validation research.

The GRE is now scored by one human rater and eRater. For the most part, ETS has been examining the accuracy of *eRater* in predicting holistic scores from human raters. Their research suggests that eRater is able to predict the scores of six raters with greater accuracy than two human raters. The question, then, is, are the eRater scores any more or less accurate than the scores provided by the two human raters typically used? If the criterion is the more raters the better, then the answer is obviously, yes. The science of psychometrics depends on the sheer magnitude of neumbers in order to statistically prove anything. A traditional direct writing assessment like holistic scoring generates a single score, technically a one item test. Because reliability is greatly improved by the number of scores, it is easy to see how subtly and quickly the question can turn to reliability. In the case of Smith's (1993) accuracy of placement, accuracy focuses on the decision and the underlying principle that all decisions are not equal. *eRater*, however, focuses on the predictive power of one set of procedures compared to another. For validity, the real question for *eRater* is whether the scores help make better decisions about students than the current procedures used by a particular college or university.

For those of us who use traditional holistic scoring procedures, the answer is likely to be that it does, because *eRater* is going to provide more stable scores than two holistic raters. However, the real test of the validity of *eRater* may lie in a comparison with procedures like Smith's that focus on the expert knowledge of teachers who determine whether the student who wrote the essay belongs in their course or the one above or below it. In this case, it is not clear that one procedure has an advantage over the other because there has never been an attempt to examine the relative value of eRater compared to the expert placement model defined by Smith.

Because the immediate question of the validity of automated scoring turns on reliability, as Huot (2002) asserts, reliability has always been the focus of the debate about writing assessment. Thus, the question of which assessment provides the best judgment of a student's placement into a writing program has still

not been answered. As various new assessments have been created (e.g., Broad, 2003; Haswell, 2001; Murphy & Underwood, 1998; Royer & Gilles, 2003), there has been a pressing need to document that these assessments promote valid and reliable educational decisions about students, teachers, and programs. Unfortunately, systematic and rigorous attention is not always given to things like consequences for various participants in the assessment.

For placement, the study of the validity of writing assessment should be focused, like Smith's, on the decision about the best course for a student to enter the writing program at a particular college. Writing exit examination validation research should be focused on a decision about a student's mastery of the curriculum, for both college and school students. Furthermore, there is little reporting of validation research in the assessment literature, in part, I suspect, because writing assessment is a field marginalized by most writing teachers and researchers. Most teachers, with good reason, fear any use of assessment, because assessment has become highly politicized by federal and state government, as well as by local school boards and administrators.

CAN COMPUTERS REPLACE ENGLISH TEACHERS?

Ultimately, one question that may cause an implicit fear is the unspoken potential for the role of automation in education as a whole, not just assessment. Does the future suggest that teachers can be replaced by computers or some evolutionary mutation of them or that one teacher via distance education technology can instruct innumerable students at various locations? One primary question I am attempting to examine is whether automated assessment should be seen as a potential threat or benefit. This fear has been the root of response to automation because automation has typically reduced the workforce in any industry. The curriculum research of the 1970s and 1980s is best summarized as an attempt to find the holy grail of education, a curriculum that is teacher proof, in the sense that the training and experience of a teacher are irrelevant to its success. The tepid results of that search are probably the reason experimental comparison of curriculums disappeared. The most valuable lesson that emerged is the importance of the teacher. Trained, experienced, and motivated teachers are the heart of successful education, despite the public furor over teachers' qualifications. Darling-Hammond and Youngs (2002) examine hundreds of studies about educational progress of various kinds of students and found that the overriding variable, more than ethnicity or income, that predicted student success was the teacher.

Many futurists, both utopian and dystopian, have seen the future filled with intelligent machines. At this stage, Anson's (2003) suggestion may be the best view, there is little that AI can offer a writing teacher. However, our real concern

should be how AI might augment the teaching of writing in the future. Explicit views of the future are not of much value, particularly because the likelihood of automation replacing some aspects of teaching writing is already evident, as we have been seeing, the continuing use of electronic technology to compliment or replace some of the work of teachers. As we have also experienced, there will be those who claim that computers allow for greater efficiency, justifying increasing the numbers of students working with individual teachers. It seems clear that computers are here to stay in English Studies, even if only as word processors to make the production of paper text easier and as communication devices to connect writers to one another for responding. We have to expect that the future will also hold some developments that can help us and some that can be hurtful. Some developments will be faddish, oversold by developers and producers of the technology, whereas others will enter our toolbox with the potential to help students learn if used properly. My answer to the problem of automated assessment is precisely the last point. Its potential suggests that it might have some value in writing classrooms, but it is not clear what that may be. Second, if it does have value, it will take continuing study understand the consequences and to establish the value through validity inquiry.

I am suggesting a stance on automated assessment that can best be characterized as carefully directed critique toward the developers of automated assessment. Because Pennsylvania has adopted automated assessment and the results of that automation will be used to determine funding for school districts, there is no question it is being used in regulatory ways. Why should we expect anything different? Assessment has been used as a gate keeper for as long as assessment has resulted in excluding some and including others in schooling.

Out-of-hand or outright rejection of automated assessment, a blanket condemnation, can only be self-serving. More importantly, we need to examine the use of automated scoring as we would any other assessment, according to the criteria of the most current theories on validating educational assessment. Arguing that theories of literacy do not justify the use of automated assessment, is similar to earlier arguments that indirect assessment does not have content validity. This argument is not going to be compelling with an educational measurement audience, not to mention policymakers and regular citizens. Furthermore, without an understanding of the common language of assessment as it is grounded in the social sciences research methodology, we will find that our righteous indignation, our hermeneutic arguments about the meaning of new types of assessment, are met by a wondering stare, at best, and a dismissive glare, at worst.

What I am arguing we do is to study automated assessment in order to explicate the potential value for teaching and learning, as well as the potential harm. The theory of the developers can itself be used as a ground for validity

arguments. However, we have to be willing to look outside our field to understand the theory of another, a theory that has clearly been at the heart of assessment practices in our culture for more than a century, and an industry that has become embedded in education in America over the last 50 years. The practices are accepted by most Americans as valid for use in education. If educators have not been successful in opposing those practices, it may be that we have not been able to understand what drives them and to be able to offer critiques that have been seen as questioning that validity.

CONCLUSION

Writing assessment in American education has two professional groups with developed bodies of theory and practice. The first group, whose primary interest is assessment, is the membership of the two professional organizations, the American Education Research Association (AERA) and the American Psychological Association (APA). They far outnumber the members of the second group, the membership of the National Council of Teachers of English and College Composition and Communication. For a number of years, APA and AERA were loosely allied through members with dual memberships. More recently, recognizing their common concerns and shared field, they began to work together. The result is a clearly defined statement of definitions and standards for test development and validation (AERA, 1999). Although the measurement community is not inherently hostile to the concerns of writing teachers, its members will be looking for the kinds of evidence articulated in the standards, applying the technology of validation research to the discussion of implementing automated scoring. Furthermore, their direct involvement with public education, as the primary source for assessment tools, lends them a strong voice in the federal, state, and local politics of assessment.

The contrasts between English Studies and educational assessment are many, running beyond concepts or methodology. The common ground is also quite large. One important point of comparison lies in the question of what constitutes important research in the two fields. In English, researchers are typically expected to demonstrate their mastery of the field in publications that are authored by a single individual. In assessment, as in most scientific fields, important research can only be conducted by a team of people, each contributing to the conceptualization and execution of the study. If it is time to examine the research methodology or social sciences as it impinges on assessment, it may also be time to explore the potential for collaborative research, not just within either a social science or humanistic tradition (see Huot, 2002, for a discussion of a unified field of writing assessment). If we continue to espouse outmoded

views of assessment, to fail to understand the complexity of validity theory, for instance, we are going to be frustrated at every turn. If for no other reason, as Sun Tsu (1994) observed, one has to know the enemy to defeat him. In this case, I hope knowing one's enemy might lead to a productive alliance.

A student of mine was attempting to articulate a complex problem for her dissertation project, one involving the value of historical study of the field. She finally told me that she recognized she was approaching the project with the wrong attitude. She said that she had forgotten a couple of the basic things she tries to teach her students: Who is the audience and what kinds of rhetorical practices are expected?

Who is our audience for our critique of automated scoring? If it is ourselves, we can continue to confront assessment developers with the challenge that their work does not conform to contemporary theories of literacy. However, when they suggest that contemporary theories of literacy are at the basis of their work, our best critique lies in a close examination of the theory, as opposed to an examination of the practice itself. Surely, well-directed critique is more successful than blanket condemnation. But, such critique emerges from the study of assessment theory, validity theory in particular. Such a critique is supported by those theories, if we take the time to use our own research skills, interpretive reading of culture icons, such as the texts of the field.

I will leave you with a story that has guided my work in the use of technology in my classroom and the suggestions that I give to others: In graduate school, I shared an apartment with a fellow student. At the time, he was working as a welder for a local company building automobile transport trailers. One day, he come in from work telling me that he had been let go. His schedule was flexible, built around his class schedule at the university. His boss had told him that the computer was not able to work with his schedule, so he had to either work full time or leave. He left and went on to accomplish some fine work in our field. However, I have adopted as a basic principle of working with computer analysts and programmers, "If your program does not do what we need it to do, you have done a poor job, go back and fix it!" The goals of people must drive the development of automation, not the automation itself. We have to find the right way to say, "Fix it!" The real trick is to get the right people to listen. As inheritors of the tradition of rhetoric, writing teachers should know more about how to speak to their audiences.

REFERENCES

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

Ajay, H. B., Tillett, P. I., & Page, E. B. (1973). Analysis of essays by computer (AEC-II). Final report to the National Center for Educational Research and Development (Project No. 80101), p. 231.

Anastasi, A. (1976). *Psychological testing* (4th ed.). Macmillan.

Anson, C. R. (2003). Responding to and assessing student writing: The uses and limits of technology. In P. Takayoshi & B. Huot (Eds.), *Teaching writing with computers* (pp. 234- 246). Houghton Mifflin.

Berlin, J. A. (1984). *Writing instruction in nineteenth-century American colleges*. Southern Illinois University.

Blakesley, D. (2003). Directed self-placement in the university. In D. Royer & R. Gilles (Eds.), *Directed self-placement: Principles and practices* (pp. 31-48). Hampton Press.

Bleich, D. (1975). *Readings and feelings: An introduction to subjective criticism*. National Council of Teachers of English.

Breland, H. M., Camp, R, Jones, R. J., Morris, M. M., & Rock, D. A. (1987). *Assessing writing skill.* College Entrance Examination Board Research Report No. 11. Educational Testing Service.

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

Burns, H. L. (1979). *Stimulating rhetorical invention in English composition through computer-assisted instruction*. Dissertation Abstracts International, DAI-A 40/70, p. 3734, January 1980, DAI Order number AAT 7928268.

Burstein, J. C. (2003). The E-rater® Scoring Engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113-121). Erlbaum.

Carmines, E. G., & Zeller, R. A. (1980). *Reliability and validity assessment*. Sage Publications.

Cooley, W. W., & Lohnes, P.R. (1976). *Evaluation research in education: Theory, principles, and practice*. Irvington.

Coombs, D. H. (1969). Review of *The Analysis of Essays by Computer*, by Ellis B. Page and Dieter H. Paulus. *Research in the Teaching of English*, *3*(2), 222-228.

Cooper, C. R., & Odell, L. (1977). *Evaluating writing: Describing, measuring, judging*. National Council of Teachers of English.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.) (pp. 443-507). American Council on Education.

Cronbach, L. J. (1988). Five perspectives on test validity argument. In H. Wainer (Ed.), *Test validity* (pp. 3-17). Erlbaum.

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

Cureton, E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621-694). American Council on Education.

Darling-Hammond, L., & Youngs, P. (2002). Defining "highly qualified teachers": What does "scientifically-based research" actually tell us? *Educational Researcher*, *31*(9), 13-25.

Elbow, P., & Yancey, K. B. (1994). On the nature of holistic scoring: An inquiry composed on email. *Assessing Writing, 1(1)*, 91-108.

Gottshalk, F. I., Swineford, F., & Coffman, W. (1966). *The measurement of writing ability.* College Entrance Examination Board Research Monograph N. 6. Educational Testing Service.

Guilford, J. P. (1954). *Psychometric methods.* New York: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests*. Wiley.

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

Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? *College English, 63(4)*, 480-499.

Huot, B. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment* (pp. 206-236). Hampton Press.

Huot, B. (1996). Computers and assessment: Understanding two technologies. *Computers and Composition, 13(2),* 231-244.

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

Johnson-Laird, P. N. (1977). *Mental models: Towards a cognitive science of language, inference, and consciousness.* Harvard University Press.

Kuhn, T. S. (1996). *Structure of scientific revolutions* (3rd ed.). University of Chicago Press.

Macrorie, K. (1969). Review of The Analysis of Essays by Computer by Ellis B. Page and Dieter H. Paulus. *Research in the Teaching of English, 3(2)*, 228-236.

Messick, S. (1989). Test validity. In R. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13- 103). American Educational Research Association; National Council of Measurement in Education.

Moss, P. (1998). The role of consequences in validity theory. *Educational measurement: Issues and practices, 17*(2), 6-12.

Murphy, S., & Underwood, T. (1998). Interrater reliability in a California middle school English/language arts portfolio assessment program. *Assessing writing, 5*(2), 201-230.

Page, E. B. (1966. The imminence of grading essays by computer. *Phi Delta Kappan,* 238-243.

Page, E. B. (1967a). Grading essays by computer: Progress report. Proceedings of the 1966 Invitational Conference on Testing (pp. 87-100). Educational Testing Service.

Page, E. B. (1967b). Statistical and linguistic strategies in the computer grading of essays. Proceedings of the Second International Conference on Computational Linguistics. Grenoble, France.

Page, E. B. (1985). Computer grading of student essays. In T. Husén & Postlethwaite (Eds.), International Encyclopedia of Educational Research (pp. 944-946). Pergamon.

Page, E. B. (1993. New computer grading of student prose, using a powerful grammar checker. [Paper presentation]. Annual meeting of the North Carolina Association for Research in Education. Greensboro, NC.

Page, E. B. (1995). Computer grading of essays: A different kind of testing? [Invited address] American Psychological Association, Divisions 5, 7, 15, 16.

Page, E. B., Fisher, G. A., & Fisher, M. A. (1968). Project Essay Grade: A FORTRAN program for statistical analysis of prose. *British journal of mathematical and statistical psychology, 21*(1), 139.

Page, E. B., Tillett, P. I., & Ajay, H. B. (1989). Computer measurement of subject-matter essay tests: Past research and future promise. Proceedings of the First Annual Meeting of the American Psychological Society, Alexandria, VA.

Pula, J., & Huot, B. (1993). A model of background influences on holistic raters. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment* (pp. 237- 265). Hampton Press.

Royer, D., & Gilles, R. (2003). *Directed self-placement: Principles and practices.* Hampton Press.

Schendel, E. (1999). Exploring the theories and consequences of self-assessment through ethical inquiry. *Assessing writing, 6*(2), 199-227.

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

Stalnaker J. M. E. (1951). The essay type of examination. In E. F. Lindquist (Ed.), *Educational Measurement* (pp. 495-532). American Council on Education.

Sun Tsu. (1994). *The art of war.* Westview.

Terman, L. M. (1916). *The measurement of intelligence.* Houghton Mifflin.

Veal, R. A., & Hudson, S. A. (1983). Direct and indirect measures for the large-scale evaluation of writing. *Research in the Teaching of English, 17*(3), 285-296.

White, E. M. (1994). *Teaching and assessing writing: Recent advances in understanding, evaluating, and improving student performance.* Jossey-Bass.

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

Williamson, M. M. (1994). The worship of efficiency: Untangling theoretical and practical consideration in writing assessment. *Assessing writing, 1*(2), 147-174.

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

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