Retrievable at http://www.tc.columbia.edu/tesolalwebjournal

Automated Essay Scoring: A Literature Review

Ian Blood[1]

Introduction

In recent decades, large-scale English language proficiency testing and testing research have seen an increased interest in constructed-response essay-writing items (Aschbacher, 1991; Powers, Burstein, Chodorow, Fowles, & Kukich, 2001; Weigle, 2002). The TOEFL iBT, for example, includes two constructed-response writing tasks, one of which is an integrative task requiring the test-taker to write in response to information delivered both aurally and in written form (Educational Testing Service, n.d.). Similarly, the IELTS academic test requires test-takers to write in response to a question that relates to a chart or graph that the test-taker must read and interpret (International English Language Testing System, n.d.). Theoretical justification for the use of such integrative, constructed-response tasks (i.e., tasks which require the test-taker to draw upon information received through several modalities in support of a communicative function) date back to at least the early 1960's. Carroll (1961, 1972) argued that tests which measure linguistic knowledge alone fail to predict the knowledge and abilities that score users are most likely to be interested in, i.e., prediction of actual *use* of language knowledge for communicative purposes in specific contexts:

> An ideal English language proficiency test should make it possible to differentiate, to the greatest possible extent, levels of performance in those dimensions of performance which are relevant to the kinds of situations in which the examinees will find themselves after being selected on the basis of the test. The validity of the test can be established not

[1] Ian Blood received his M.A. in Applied Linguistics from Teachers College, Columbia University, and is a current Ed.M. student in the same program. He works as a research assistant at Educational Testing Service. His current research interests include classroom-based assessment and reading assessment validity. Correspondence should be sent to iab2111@tc.columbia.edu

> solely on the basis of whether it appears to involve a good sample of English language but more on the basis of whether it predicts success in the learning tasks and social situations to which the examinees will be exposed. (p. 319)

Moreover, Canale and Swain (1980), in their elaboration on the notion of communicative competence, explained that demonstration of linguistic mastery is a necessary but not sufficient condition for inferring communicative language ability on the part of an individual, and that students of foreign languages must be given opportunities "to respond to genuine communicative needs in realistic second language situations … not only with respect to classroom activities, but to testing as well" (p. 27).

Bachman and Palmer (1996) echoed the concern that most language testing purposes are such that test constructs should define communicative language *ability* and not simply linguistic knowledge. In their discussion of the qualities of language tests that enhance test usefulness, they included authenticity of test tasks as a major contributing element. Authenticity in this context is defined as "the degree of correspondence of the characteristics of a given language test task to the features of a TLU task" (p. 23). If test tasks closely resemble the real-world language use situations that are of interest to stake-holders, then it is assumed that this will enhance the predictive power of the tasks, and, therefore, the overall validity and usefulness of the test.

When the target language use domain of1` interest to test-developers is that of the writing activities of university students, it has often been seen as a prima facie minimum requirement for authenticity that the test tasks involve the production of an actual written response that requires examinees to integrate knowledge in service of a communicative task in the way that might be called for when writing an academic essay. Popham (1978) does not mince words on the topic:

> As a measure of certain kinds of complex learning outcomes, the essay item is unchallenged. Since this kind of item sometimes requires a student to put together ideas and express them in original ways, there is no way of simulating that kind of requirement in a selected-response item or, for that matter, even in a short-answer item. (p. 67)

However, constructed-response writing tasks have both advantages and disadvantages. Unlike multiple-choice items which have a single criterion for correctness, experts often disagree on how to operationalize and score the set of qualities that define excellent writing. While the fact that constructed-response essay items require students to generate samples of normative language (rather than simply selecting them) may make such items a more proximal measure of communicative writing ability, the process of scoring essay items is quite complex. Human raters must be hired and trained to score each of the examinee essays. The additional time required for this process means that test scores cannot be reported to examinees as quickly as would be possible for machine-scored multiple-choice items (Livingston, 2009). Moreover, the costs associated with this process are passed on to the examinees themselves in the form of testing fees. In addition, the use of human raters introduces a new challenge to maintaining the reliability and construct validity of test scores, as raters are bound to differ in their perceptions of candidate performances and their tendencies towards leniency and severity. Raters may also have unconscious biases that are not immediately amenable to correction through training (Kondo-Brown, 2002).

Perhaps it is for the above reasons that the testing industry was initially slow to include constructed-response tasks on large-scale, high-stakes assessments. It was not until 1986 that Educational Testing Service (ETS) began to offer the Test of Written English, the first performance-based examination of second language writing ability to be offered by the

organization. Prior to this time, writing ability had been measured indirectly on the TOEFL with multiple-choice items focusing on "knowledge of structure and written expression" (Stansfield, 1986, p. 4). Also for the reasons mentioned above, the advent of computer technology has led to multiple attempts to create automatic scoring applications that might allow for cheaper, more reliable, and in the view of some proponents, even more accurate scoring of test-taker performances. The pages that follow will provide a discussion of some of the challenges associated with the performance-based assessment of writing as well as a critical review of the automated essay scoring applications that have been developed as an alternative to or supplement for human ratings of writing performance.

Challenges Associated with Essay Scoring

A basic requirement of ethical language testing is that test developers provide evidence in the form of research findings for the interpretive argument that underlies specific claims regarding the inferences that may be justified by test scores. When assessing writing ability on the basis of a test taker's written performance, construct-irrelevant and unreliable variance in the ratings are a chief concern (Kondo-Brown, 2002; McNamara, 1996; Weigle, 2002). Such variance may occur when raters differ in their perceptions and preconceived notions regarding good writing and good writers. If individual raters lack the ability to give consistent ratings (intra-rater reliability) and to rate in a normative way (i.e., in the way that is intended by the test designer, consistent with the descriptions in the rating scale, and in agreement with other raters [inter-rater reliability]), the validity of inferences based on the scores of writing performance assessments will be diminished. In order to maximize the intra-rater and inter-rater reliability of ratings given by human raters, the usual practice is to have at least two raters assign either holistic or component scores to written performances with the aid of a rating scale and following

some kind of rater training. In actual practice, however, rater training often consists of a single norming session in which raters practice giving ratings and compare their own ratings to those of other raters (Kim, 2010). Achieving consistency of rating and normative rating behaviors on the part of essay raters requires a significant investment on the part of both testing programs and the raters themselves. Testing programs must train raters and pay them for their time, while the raters themselves must spend long hours rating multiple essays at a sitting. Kim (2010) found that repeated rater training and feedback sessions were necessary to achieve internal consistency and normative rating behavior on the part of even experienced raters. Even large testing companies may struggle to allocate the resources necessary to ensure that raters receive adequate training. Several studies have highlighted the challenge of assigning reliable and construct-valid scores when using human raters to score writing performances (Kondo-Brown, 2002; Shi, 2001; Shohamy, Gordon, & Kraemer, 1992).

Shohamy *et al.* (1992) examined intra-rater and inter-rater reliability for a diverse group of raters who were divided into a group that received rater training and a group that did not. An encouraging finding of the study was that, compared to rater background, rater training had a much larger and positive effect on both intra-rater and inter-rater reliability. However, the authors note that intensive, repeated rater training including ample opportunity for discussion and feedback is needed to achieve the most desirable levels of reliability. Training of this type may not be feasible for large testing companies employing thousands of raters in remote locations.

Kondo-Brown's (2002) study is a good illustration of a different challenge to the reliability of human ratings. The study examined the severity of ratings given to Japanese L2 writing performances by a group of raters who were homogeneous with respect to background

and native language. The raters in the study were all native speakers of Japanese, taught in the same university, and held advanced degrees in language-related fields. While high inter-rater correlation coefficients were observed for the three raters in the study, the raters nevertheless displayed significant individual differences in severity with respect to particular aspects of writing performances. In particular, significantly biased ratings were more common for candidates with very high or low ability, suggesting that despite rater training the raters were not able to be consistent in their application of rating criteria to examinees at all ability levels. Kondo-Brown (2002) concludes that while rater training may improve inter-rater reliability and the internal consistency of ratings, it may not have much impact on other undesirable rater characteristics such as increased or decreased severity when rating certain examinees, items, or for certain areas of the rating scale.

Similarly, Shi (2001) examined holistic ratings given by native and non-native English-speaking raters of EFL writing in a Chinese university and found that, while the two groups did not differ significantly in the scores they gave, the raters differed greatly in the justifications that they gave for their ratings. Self-report data suggested that different raters weighted different essay characteristics more heavily while arriving at a holistic score. Native speakers tended to focus more on content and language while non-native speakers focused on organization and length. Therefore, although they may have given similar scores to an essay, the raters in Shi's study gave these scores for very different reasons, suggesting that they did not share a common understanding of the test construct. Shi notes that if students were to receive feedback from raters such as the ones in her study, they would likely be confused by contradictory messages. While it is often assumed in the field of language assessment that human ratings, once reliable, are inherently valid, Shi concluded that the findings underline the lack of a one-to-one

correspondence between the reliability and construct-validity of human ratings and shed light upon the need for the development of rating procedures that promote more construct-valid ratings, such as the use of analytic rubrics that encourage more thorough and balanced attention to the construct.

Automated Essay Scoring Systems

Project Essay Grade

Just several decades before the arrival of automatic essay scoring technology, the release of the IBM 805 automatic selected-response scoring machine in 1938 (IBM, n.d.) heralded a period of great expansion in the use of multiple-choice tests. From the beginning, this development was seen by some as an advance in the objectivity, economy, and speed of standardized testing, and by others as a restrictive, inauthentic, and ultimately unfair means of measuring examinee knowledge and abilities (Martinez & Bennett, 1992). The earliest example in the literature of the development of an automatic essay scoring tool is Project Essay Grade (PEG), developed by Ellis Page in the mid 1960's. Just as with the arrival of the IBM 805, PEG was met with both excitement and skepticism.

Page himself was a believer in the educational value of writing activities, and he saw automatic essay scoring's potential to lighten the workload of teachers—and its resulting potential to promote the increased use of writing assignments in instruction—as a major benefit (Page, 1968). An assumption upon which PEG was developed is the hypothesized divisibility of essay *content* (the information contained in the essay) and essay *style* (the way in which the information is conveyed). PEG was designed to assign scores based solely on the style of candidate essays, i.e., PEG attended only to linguistic surface features without regard to the normativity of the candidates' ideas. Even the candidate's syntax could not be directly evaluated by PEG as natural language processing (NLP) capabilities were still nascent.

Page (1968) also categorized two types of variables that may be of interest in scoring an essay: what he termed *proxes* and *trins*. A prox is a variable that a computer can be programmed to recognize in a text (i.e., an indicator variable), while a trin is a variable of the type that an expert human rater would be concerned with (e.g., a latent variable). An example of a prox would be the total number of words in an essay, which may be assumed to have a positive relationship with the trin variable *elaboration* (I will henceforth refer to proxes and trins as indicators and latent variables). However, Page was not concerned with the latent variables that are commonly discussed in writing theory, such as *organization* or *meaning*. Rather, the sole criterion that PEG was designed to predict was an average single-number holistic score given by a group of human raters. Moreover, Page was not concerned with whether the surface features that he chose as indicators had the appearance (or lack thereof) of validity with respect to theories of writing. His sole interest was in finding machine-interpretable features that would both strongly and reliably predict the human score criterion.

To translate this idea into a research design, Page collected a sample of 276 high school essays and had at least four expert raters assign holistic scores to each of them. These scores were treated as the criterion variable in a multiple-regression equation. Page and his colleagues then examined the essays to hypothesize observable indicators for these variables. Thirty indicators were settled upon including more intuitive features (presence of a title and number of paragraphs) and less intuitive features (number of apostrophes in the essay and standard deviation of word length). Having identified the indicator variables, Page used the regression equation to assign weights to each indicator based on its prediction of the criterion. The most heavily weighted indicator identified by Page was the length of the essay in words, followed by

the standard deviation of word length[2]. After these and other indicators were coded into PEG, the program was used to rate the original sample of essays, and the ratings generated were compared with the average holistic ratings of the human rater group. Page (1968) reported a correlation coefficient of .71 with the expert raters.

A central concern, however, was how well the indicators and associated weights used by PEG would generalize to *other* rating situations. Therefore, Page also examined how well the program, having been trained on the original sample of essays, would compare to human raters when rating a *new* sample of essays. For this purpose, Page calculated an inter-rater correlation matrix for PEG and four human raters based on their ratings of a new sample. Page reported that PEG's correlation with the human raters was close enough to be indistinguishable. Later versions of PEG achieved correlation coefficients of up to .77 with human ratings, a figure higher than the average correlations observed between individual human raters and the entire human rater group, leading Page to conclude that PEG performed more reliably than individual human raters (Page, 1968).

E-Rater

Over 40 years after Page's first attempt, many automatic essay scoring tools still resemble PEG in an important way. While the indicators attended to may have become more sophisticated, the technology still approximates the *product* (scores) rather than the *process* (rating behaviors) of human raters. One example, an automatic essay scoring tool that is in operational use for high-stakes testing purposes, is E-rater; it was developed by ETS in the mid-1990's (Burstein, Braden-Harder, Chodorow, Hua, Kaplan, Kukich, Lu, Nolan, Rock, & Wolff, 1998). E-Rater, like PEG, predicts human ratings with a multiple-regression equation. However,

[2] Surprisingly, essay length has continued to be demonstrated to strongly predict human raters' holistic scoring of essays (Attali & Burstein, 2006).

one distinction between E-Rater and PEG is E-Rater's use of natural language processing (NLP) technology, allowing for the inclusion of grammatical accuracy as an indicator in the regression equation. Nevertheless, the earliest version of E-Rater had more similarities with PEG than differences as it made use of a large number of mostly non-intuitive surface features with essay length the most heavily weighted amongst them. E-Rater V.2 was released in 2006, partly as a response to criticism of E-Rater V.1's disproportional attention to essay length in assigning scores. The new version added a measure of the degree of overlap between words in candidate essays and other previously scored essays that have received a top score for the same prompt. This feature is considered to be an indicator of topicality (Ben-Simon & Bennett, 2007). E-Rater V.2's designers reduced the number of indicators that it attends to 12 and grouped them under the more intuitive headings of *grammar, usage, mechanics, & style*; *organization & development*; *topical analysis*; *word complexity*; and *essay length*. This relatively small number of indicators represents an attempt to make the program's operation more reflective of a writing construct and more readily recognizable to score users as related to what writers do. The model is also intended to decrease the overall weight of essay length in predicting the final score, and thus decrease the likelihood that "bad faith" essays can achieve high scores (Attali & Burstein, 2006).

E-Rater is currently used by ETS for both high- and low-stakes testing situations, including the scoring of the Issue and Argument essay items on the Graduate Record Exam (GRE), the independent writing section of the internet-based version of the Test of English as a Foreign Language (TOEFL), and several online writing practice tools. ETS claims E-Rater has resulted in higher quality scores and faster delivery of test results to candidates (ETS, 2010). Attali and Burstein (2006) performed a multi-method analysis of E-Rater's validity comparing

scores given by E-Rater, single human raters, and average scores from two raters to different essays. They reported a correlation of .97 between E-Rater and the average human score, which was well above the minimum expected correlation between individual human raters.

Latent Semantic Analysis

An alternative to what Ben-Simon and Bennett (2007, p. 16) refer to as the "brute empirical" (surface feature) approach to automatic essay scoring applied in PEG and E-Rater is the latent semantic analysis-based approach (LSA). LSA is described by its developers as "a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text" (Landauer, Foltz, & Laham, 1998, p. 2). LSA was developed not only as a practical tool for applications, such as meaningful corpus searching and the scoring of essays, but also "as a model of the computational processes and representations underlying substantial portions of the acquisition and utilization of knowledge" (p. 3). Unlike PEG, which focuses entirely on incidental surface features, or E-Rater, which combines incidental surface features with the more intuitive, such as vocabulary and grammatical accuracy, LSA ignores superficial surface features and word order entirely, attending instead to the statistical relationships in an examinee essay between meaningful units (i.e., words, sentences, paragraphs, and texts). Understanding the processes that underlie LSA is daunting for the mathematically uninitiated, but its basic premise is that the meaning of a word, sentence, or text, and the concepts embodied by each, are closely related to the *conceptual* contexts in which each occurs. Therefore, by measuring the collocations of meaningful units of text, a computer program can 'learn' what knowledge or concepts are contained in a piece of writing.

Foltz, Landauer, and Laham (1999) developed an LSA-based automatic essay scoring tool, Intelligent Essay Assessor (IEA), designed to attend to essay *content* rather than *style*. Unlike PEG and E-Rater, IEA trains not on a sample of essays scored by human raters, but on domain-representative text (e.g., actual essays written by college students or course textbooks). IEA, therefore, is used to determine the degree of congruence between an essay of known quality and a candidate essay. Since LSA represents semantic information as likely collocations between both words and similar words/groups of words in larger units of text (and not simply by word matching), Foltz *et al.* (1999) claim that it can recognize essays that are similar in content even if they differ markedly in vocabulary, grammar, and style. This means that IEA might be used to offer substantive feedback on the content of examinee responses. In order to avoid inaccurate scoring of "unique" essays that may not be appropriately evaluated by IEA's algorithm, such essays are flagged as anomalous and rated by a human. Currently, IEA will flag essays that it deems highly creative, off topic, in violation of standard formatting, or too similar to another known essay (and thus a possible case of plagiarism).

To evaluate the accuracy of IEA in an actual scoring situation, Foltz *et al.* (1999) scored a sample of over 600 opinion and argument GMAT essays. IEA achieved correlations of .86 with the human graders for both the opinion and argument essays while the ETS raters correlated with each other at .87 and .86 for the two opinion and argument essays, respectively. Foltz *et al.* (1999) also reported on other evaluations of IEA for different content areas including psychology, biology, and history at the middle school, high school, undergraduate, and graduate levels. In each case, IEA's observed reliability was comparable to inter-rater reliability and within generally accepted ranges for the testing purpose. Foltz *et al.* (1999) also reported that other

LSA-based tools have scored comparably to humans on content-based assessments, such as the synonym portion of the TOEFL and a multiple-choice introductory psychology exam.

As an additional, qualitative line of inquiry into IEA's usefulness, Foltz *et al.* (1999) recruited a university psycholinguistics course to use IEA for essay scoring over the course of two years. IEA was trained for this purpose on a sample of passages from the course textbook. To verify reliability before operational use of IEA, the researchers graded sample essays from previous semesters with the program and observed a correlation of .80 with the average human scores. Students in the course were able to submit essays online and receive instant estimated grades accompanied by feedback and suggestions for subtopics that may be missing. The students were encouraged to use IEA as a tool for revising their essays iteratively until they were satisfied that their essays were ready to be submitted to the professor for final grading. In a survey at the conclusion of the study, 98% of the students reported that they would definitely or probably use IEA if it were available for their other classes. Foltz *et al.* (1999) suggest that IEA has potential to impact assessment at all levels, including as a supplement or replacement for human raters in standardized testing, as an aid and objective check for classroom teachers, and as a content feedback device for students.

Text Categorization

Another approach to automatic scoring is the application of text categorization technology. Williams (2001) explains that the text categorization method uses Bayesian independent classifiers attending to word occurrence "to assign probabilities to documents estimating the likelihood that they belong to a specified category of documents" (p. 5). When applied to essay scoring, previously scored essays are divided into categories of quality, and the algorithm is used to predict which quality category a newly scored essay would most likely be

assigned to by a human rater. Larkey (1998) developed an automatic essay scoring tool that integrated text categorization technology with the identification of 12 linguistic surface features to predict human scores in a regression equation. In order to evaluate the tool's performance, it was used to assign scores to samples of at least 200 essays on the topics of social studies, physics, law, and two general questions intended to measure the writing ability of college students seeking graduate school admission. The algorithm performed better for social studies and law than it did for physics and better on the two general questions than it did for any of the specific content areas. However, all correlations with human graders were high, in the .70s and .80s. Larkey (1998) also examined the frequency with which her scoring tool would agree exactly with human raters, and the frequency with which it would differ from human raters by only one point on the rating scale. The tool agreed with raters exactly 50 to 65% of the time, and differed by one or less 90 to100% of the time.

Validity, Strengths, and Weaknesses

Kane (2006) regards validity in educational measurement as a concept that must be supported by two interrelated arguments: an *interpretive argument* and a *validation argument*. The interpretive argument "specifies the proposed interpretations and uses of test results by laying out a network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performances," (p. 23) and includes scoring, generalization, extrapolation, and interpretation steps. The validation argument serves to evaluate and challenge the interpretive argument by providing evidence related to individual inferences and assumptions. This conception builds off of Messick's (1989) framework (Table 1) which represents the facets of validity as a progressive matrix in which *test interpretation* and *test use* are supported by *evidential* and *consequential* bases.

Table 1
Messick's (1989) Facets of Validity

	TEST INTERPRETATION	TEST USE
Evidential Basis	Construct validity	Construct validity +Relevance/utility
Consequential Basis	Value implications	Social consequences

The framework is progressive in that construct validity, represented in the matrix as test interpretation supported by evidence, is a necessary but not sufficient condition for each neighboring cell. For a claim of validity to be made, then, evidence must support the appropriateness of test interpretation and test use; it must also be demonstrated that the value implications implicated in test interpretation and the individual and social consequences of test use are *beneficial* to the stake-holders.

It is important to note that Kane's (2006) argument-based approach is not static and is not a simple macrostructure of validation studies that should be applied to every assessment. Rather, an interpretive argument for validity entails the elaboration of a basic framework of inferences and assumptions undergirding use of the assessment that will necessarily differ from one context to another. Moreover, the validation argument, the research activities that evaluate the interpretive argument, will also necessarily depend upon the testing purpose. Therefore, to speak of the validity of a test or of an automatic essay scoring application, is to miss the point. We cannot evaluate the validity of automatic scoring tools in the abstract. Rather, each particular claim regarding the use of automatic scoring for a particular purpose must be evaluated on its own merits. While studies that produce evidence of a general nature regarding the capabilities of a scoring tool are useful insofar as they inform an understanding of how the tool *might* be of use to the assessment field, these studies do not tell the whole story.

As an exercise in contextualized instantiation of an interpretive argument, Kane (2006) describes a minimal set of inferences and assumptions that would need to be made for a placement testing system. It will be useful, for the purposes of this review, to consider how the use of automatic scoring systems would impact such an argument.

Automatic scoring tools directly impact the first of the four major inferences laid out by Kane: *scoring* or the correspondence between observed performance and observed score. In the context of proficiency testing, Kane suggests that the inference stands on two assumptions: "the scoring rule is appropriate" and "the scoring rule is applied accurately and consistently" (p. 24). Evaluative studies of automatic essay scoring tools have presented evidence related to both of these assumptions. To justify the assumption that scoring rules are appropriate, researchers have argued either that scores given by automatic essay scoring tools closely reflect the *product* of human ratings (e.g., Attali & Burstein, 2006; Page, 1968) or that they employ similar *processes* as those employed by human raters (e.g., Foltz *et al.*, 1999). The evidence that has been brought to bear most frequently on this question is correlation between machine and human scores. The matter of whether or not automatic essay scoring tools apply scoring rules accurately and consistently has not been a matter of debate as internal consistency is an inherent trait of such systems.

Many of automatic essay scoring's most vocal critics have failed to contextualize their criticism within an interpretive argument for the validity of a specific testing context, thus relegating their objections to the realm of the abstract. Dreschel (1999), for example, objects to automatic essay scoring on the grounds that it "ignores composition research" and "reduce[s] the assessment of complex cognitive processes such as writing to a mere evaluation of grammar and mechanics" (p. 1). The Conference on College Communication and Composition (CCCC, 2004)

argue that "writing-to-a-machine violates the essentially social nature of writing: we write to others for social purposes" (p. 1). Cheville (2004) states that "we know that the standards of correctness that constitute the diagnostic domains of a program like Criterion[3] are arbitrary. They reflect the selective attention of a cadre of computational linguists whose technical sophistication is oriented not solely to what language can do but rather to what machines can do with language" (p. 50). Putting aside the fact that human judges themselves often struggle to agree upon standards of correctness (e.g., Shi,2001), the preceding objections to automatic scoring are interesting for the philosophical stance that they reveal. I would argue that the position embodied by these quotations can be described as an *absolutist* approach to validity: *validity can only be argued from the justification of scores with reference to theory "X"*. The position embodied by Kane (2006) and Messick (1998) might be described as a *relativist*, or ends-justify-the-means approach: *The importance of theory "X" to scoring is a function of how well theory "result" in scores that accurately predict ability in the target domain.* To illustrate this with an example, one might imagine two idealized testing situations: a large-scale, high-stakes proficiency test with an essay component, and a low-stakes, classroom-based, formative assessment of writing ability. In the case of the proficiency test, highly reliable and accurate rank-ordering of candidates with respect to the criterion of academic writing ability is the overarching goal. Definition of the construct of academic writing ability for this testing purpose should be informed by theories of writing and evidence of the demands of university life (e.g., a review of assignments, classroom-based tests, student writing identified stakeholders as exemplary, and professor expectations in a sample of universities). In the case of the classroom-based assessment, however, the overarching goal is not rank-ordering, but rather the provision of relevant and detailed feedback to learners on their writing ability. In this case the construct may need to be operationalized in more precise

[3] *Criterion* is an ETS automatic scoring platform that is targeted to the k-12 market.

terms with respect to hypothetical subcomponents or sub-skills. The question of the potential usefulness of automatic scoring now becomes more focused: To what degree can automatic scoring tools provide scores that justify the *specific inferences* that test users need to make regarding ability in the construct? If the scoring tool in question is trained on surface features of candidate responses in order to predict the average holistic-scale rating of human raters with high reliability, we are likely to conclude that it is less appropriate for testing situations in which substantive feedback is highly valued. By the same token, we may decide that the speed, reliability, and low cost of the tool make it an attractive candidate for use scoring the proficiency test as providing feedback to the candidates, while desirable, is not a major intended outcome of such a test. The debate surrounding automatic essay scoring will be more fruitful if it focuses on real or at least hypothetical instantiations of automatic scoring in the context of a known testing purpose.

Among the more compelling criticisms of automatic essay scoring is the concern that awareness of automatic scoring on the part of students and teachers will result in negative washback (Alderson & Wall, 1993) effects on instruction and study habits. As the argument goes, if students become aware that essays will be machine scored, they may begin to write to the *machine* rather than a human audience, attending only to those features (such as essay length and vocabulary choice) that most strongly affect their scores. Likewise, the concern has been raised that teachers, under pressure from administrators to raise the test scores of their students, may begin to encourage this non-productive behavior in their learners. While Dreschel (1999), Cheville (2004) and CCCC (2004) articulate these concerns in a compelling manner, they do not provide convincing evidence that students and teachers actually do react this way when exposed to automatic scoring. Qualitative evidence of the instructional practices and preparation methods

employed by teachers and students who encounter automatic scoring would need to be gathered to mount an argument for the reality of negative washback effects.

Consequence studies should also be undertaken to evaluate the claims of automatic scoring tool designers (e.g., Foltz *et al.*, 1999; Page, 1968) that such tools can have *positive* washback effects in classrooms, such as lessening teacher workload, encouraging more writing activities, and providing more opportunities for students to receive useful feedback on writing. Foltz *et al.* (1999) stand out among the studies reviewed here for its inclusion of correlational data as well as qualitative observational data on assessment use and user perceptions and for its examination of an instantiation of the IEA in actual use.

Another plausible disadvantage of automatic essay scoring tools is their vulnerability to manipulation by malicious examinees. Especially in the case of rating tools which weight one or two features, such as length, very highly, there would seem to be a potential for test-takers to beat the system and achieve a high score by simply copying words onto the page without concern for the meaning that is conveyed. While many automatic essay scoring tools in use today incorporate features to flag essays that the scoring engine is likely to score inadequately, a study by Powers *et al.* (2001) revealed that concerns about how essays written in bad faith would be graded by E-Rater V.1 were not entirely unfounded. Powers *et al.* (2001) invited writing experts to test E-Rater by composing essays with characteristics that they thought may result in scores that are not accurate predictions of what an expert human rater would assign. The experts received detailed information on the specific surface features that E-Rater attends to when scoring, and on the basis of this information they each composed two essays—one designed to receive an artificially high score and the other designed to receive an artificially low score. The essays were scored using E-Rater, and the scores were compared with those of human raters.

When the results were tabulated, 26 out of 30 essays were correctly hypothesized by the experts to be scored higher by E-Rater than by human judges, and 10 of 24 essays were correctly hypothesized to be rated lower by E-Rater. The largest discrepancy between E-Rater and the human raters was for an essay that was written by a computational linguistics expert and simply contained 37 copies of exactly the same paragraph, resulting in a very long essay. Powers *et al.* (2001) concluded that essays written in bad faith certainly can defeat E-Rater's ability to accurately predict human scores, and that E-Rater is, therefore, not ready to be used as the sole means of scoring essays in high-stakes assessments. However, this does not mean it would have no usefulness as a second rater, or even as the sole rater in testing contexts where bad faith essays would be unlikely to occur (e.g., in online practice applications).

Conclusions and Future Directions

The limitations of automatic essay scoring restrict the applications for which it is appropriate, and significant limitations exist in the areas of vulnerability to bad faith essays and inability on the part of most systems to provide rich feedback. However, the possible benefits, including high reliability, lower testing costs, faster score reporting, and the potential to lighten teacher workloads allowing for *more* student writing assignments, make automatic scoring an area that deserves continued research. An exciting possibility is the development of hybridized systems combining the capabilities of product-oriented (e.g., E-Rater) and process-oriented (e.g., LSA) technologies.

In order to substantiate positive and negative washback claims, research in specific use contexts needs to be performed. In view of the fact that impressive correlations with human holistic ratings have been observed for all of the tools reviewed for the present study, the

fundamental issue is not whether automatic scoring tools are viable, but which tool might be viable for what purpose. Messick's (1989) validity framework

> forestalls undue reliance on selected forms of evidence …. highlights the important though subsidiary role of specific content- and criterion-related evidence in support of construct validity in testing applications, and … formally brings consideration of value implications and social consequences into the validity framework. (p. 20)

In this spirit, qualitative as well as quantitative studies should be undertaken to evaluate automatic scoring technology, its use in diverse testing applications, and the *impact* of its use on learners.

References

Alderson, J. C., & Wall, D. Does washback exist? *Applied Linguistics, 14,* 115–129.

Aschbacher, P. R. (1991). Performance assessment: state activity, interest, and concerns. *Applied Measurement in Education, 4,* 275–288.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *The Journal of Technology, Learning, and Assessment, 4*(3), 1–31.

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice*. Oxford, UK: Oxford University Press.

Ben-Simon, A., & Bennett, R.E. (2007). Toward More Substantively Meaningful Automated Essay Scoring. *Journal of Technology, Learning, and Assessment, 6*(1), 1–47.

Burstein, J., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Kukich, K., Lu, C., Nolan, J., Rock, D., & Wolff, S. (1998). *Computer analysis of essay content for automated score*

prediction: A prototype automated scoring system for GMAT analytical writing assessment essays. ETS Research Report No. 98-15. Princeton, NJ: Educational Testing Service.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1–47.

Carroll, J. B. (1961/1972). Fundamental considerations in testing for English language proficiency of foreign students. Reprinted in Allen, H. B., & Campbell R. N. (Eds.), *Teaching English as a Second Language: A Book of Readings* (2nd ed.). NY: McGraw Hill.

Cheville, J. 2004: Automated scoring technologies and the rising influence of error. *The English Journal, 93*(4), 47–52.

Conference on College Composition and Communication. (2004). *CCCC Position Statement on Teaching, Learning, and Assessing Writing in Digital Environments*. Retrieved from http://www.ncte.org/cccc/resources/positions/digitalenvironments.

Dreschel, J. (1999). Writing into silence: Losing voice with writing assessment technology. *Teaching English in the Two Year College, 26,* 380–387.

Educational Testing Service (ETS). (2010). *ETS automated scoring and NLP technologies*. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Home/pdf/AutomatedScoring.pdf

Educational Testing Service (ETS). (n.d.) *TOEFL tour*. Retrieved January 17th, 2011 from http://www.ets.org/Media/Tests/TOEFL/tour/highrez/start-web_content.html

Foltz, P. W., Landauer, T. K., & Laham, D. (1999). Automated essay scoring: Applications to educational technology. In Proceedings of EdMedia '99. Retrieved from http://www.psych.nmsu.edu/~pfoltz/reprints/Edmedia99.html

International Business Machines (IBM). (n.d.). *They also served: An album of IBM special products (vol. 1): IBM 805 Test Scoring Machine*. Retrieved from http://www-03.ibm.com/ibm/history/exhibits/specialprod1/specialprod1_9.html

International English Language Testing System. (n.d.). *Academic writing sample*. Retrieved January 17th, 2011 from http://www.ielts.org/test_takers_information/sample_test_materials/academic_writing_sample.aspx

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kim, H. J. (2010). *Investigating raters' development of rating ability on a second language speaking assessment*. Unpublished doctoral dissertation, Teachers College, Columbia University.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19,* 3–31.

Landauer, T. K., Foltz, P. W., & Laham, D. L. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25,* 259–284.

Larkey, Automatic essay grading using text categorization techniques. Paper presented at 21st International Conference of the Association for Computing Machinery-Special Interest Group on Information Retrieval (ACM-SIGIR), Melbourne, Australia. Retrieved from http://ciir.cs.umass.edu/pubfiles/ir-121.pdf.

Livingston, S. A. (2009). Constructed-response test questions: Why we use them; how we score them. *Educational Testing Service R & D Connections, 11,* 1–8. Retrieved from http://www.ets.org/Media/Research/pdf/RD_Connections11.pdf.

Martinez, M. E., & Bennett, R. E. (1992). A review of automatically scorable constructed-response item types for large-scale assessment. *Applied Measurement in Education, 5,* 151–169.

McNamara, T. (1996). *Measuring second language performance*. NY: Addison Wesley Longman Limited.

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

Page, E. B. (1968). The use of the computer in analyzing student essays. *International Review of Education, 14,* 210–225.

Popham, W. J. (1978). *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2001). Stumping e-rater: Challenging the validity of automated essay scoring (GRE Board Professional Rep. No. 98–08bP, ETS Research Rep. No. 01–03). Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR-01-03-Powers.pdf.

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing, 18,* 303–325.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal, 76,* 27–33.

Stansfield, C. W. (1986). A History of the Test of Written English: The Developmental Year. Paper presented at an International Invitational Conference on Research in Language Testing (Kiryat Anavim, Israel, May 10–13, 1986). Princeton, NJ: Educational Testing Service.

Weigle, S. C. (2002). *Assessing writing.* Cambridge, UK: Cambridge University Press.

Williams, R. (2001). Automated essay grading: An evaluation of four conceptual models. In A. Herrmann and M. M. Kulski (Eds.), *Expanding Horizons in Teaching and Learning. Proceedings of the 10th Annual Teaching Learning Forum, 7–9 February 2001*. Perth, Australia: Curtin University of Technology. Retrieved from http://lsn.curtin.edu.au/tlf/tlf2001/williams.html.