CHAPTER 7.

CONSTRUCT VALIDITY, LENGTH, SCORE, AND TIME IN HOLISTICALLY GRADED WRITING ASSESSMENTS: THE CASE AGAINST AUTOMATED ESSAY SCORING (AES)

Les Perelman
Massachusetts Institute of Technology

Automated Essay Scoring (AES), the use of computers to evaluate student writing, first appeared in 1966 with Project Essay Grade (Page, 1994). Since 1990, the three major products have been Vantage Technologies' Intellimetric, Pearson's Intelligent Essay Assessor, and the Educational Testing Service's e-rater. Advocates of Automated Essay Scoring originally justified the efficacy of their various algorithms by the ability of AES to replicate closely the scores of human graders. This concurrent validity proved, however, to be insufficient, because as Attali & Burstein note, "In the case of AES, the significance of comparable single-essay agreement rates should be evaluated against the common finding that the simplest form of automated scoring which considers only essay length could yield agreement rates that are almost as good as human rates. Clearly, such a system is not valid" (2006, p. 5). The various AES systems then developed constructs that their creators claimed, could make their assessments more valid and reliable than human graders.

This chapter argues that although the whole enterprise of automated essay scoring claims various kinds of construct validity, the measures it employs substantially fail to represent any reasonable real-world construct of writing ability. (The term *validity i*n psychological testing refers to the ability of assessment scale or instrument to measure what it claims to be measuring. The term, *construct validity,* refers to an assessment instrument's ability to measure a theorized scientific construct that cannot be directly measured, such as intelligence, creativity, critical thinking, or writing ability.) The metrics employed by AES are not relevant to effective writing in the twenty-first century and, in many cases, detrimental to it. Its main success has been in producing correlations with

DOI: https://doi.org/10.37514/PER-B.2012.0452.2.07

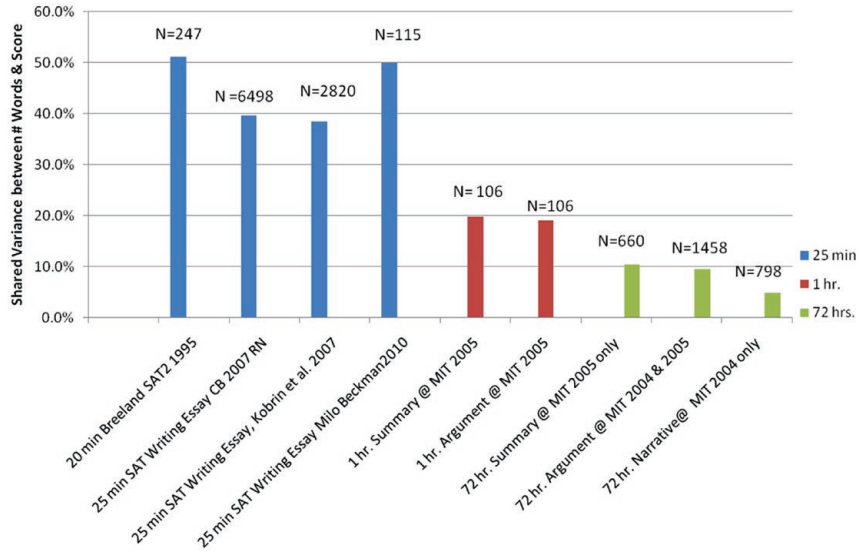

Figure 1. Shared Variance between Holistic Score and Length as a Function of Time Allowed

human grades based almost entirely on length of essays. More importantly, the importance of length in ranking essays is almost entirely an artifact of the type of artificial assessment used in most mass market writing assessments be they graded by humans or machines, the very short timed impromptu.

THE TIMED IMPROMPTU

Although White (1995) has made a case for the timed-impromptu for certain assessment decisions, it is a genre of writing that has no real analogue in real human communication and therefore is invalid as a measure. Indeed, the timed impromptu exists in no activity system except for mass-market writing assessments and education geared towards mass-market writing assessments. Writing on demand occurs in numerous situations including the traditional college essay examination. Students study for examinations to anticipate the kind of questions they will be asked and the types of information and arguments they will be required to provide. In other contexts, as well, a request for a quick written response always assumes that the writer has prior knowledge of the topic. A supervisor may ask an employee to comment on some project he or she is working on and may even want a written answer within thirty minutes, but will never ask for a response to the type of general questions that

populate mass-market writing assessments. A boss does not send an email to a subordinate stating, "'Failure is necessary for success.' Send me a well organized response to this statement in 25 minutes." People do not write on general topics on demand to no one.

In the early days of writing exams for admission or placement to American colleges and universities, the essay questions were always based on a list of set texts, almost always literary. The English Composition essay of 1874 entrance examination for Harvard College, for example, was based a reading list that included three plays of Shakespeare, and novels by Goldsmith and Scott (Elliot, 2005).

In the early twentieth century, psychologists such as Carl Brigham, the Secretary of the College Board and the subsequent developer of the Scholastic Aptitude Test, moved away from what Brigham classified as Restrictive Examinations based on specific knowledge toward what he classified as Comprehensive Examinations in English. These examinations had more open-ended questions than the earlier Restrictive Examinations and more closely resembled the kind of open-ended questions that exist now in the timed-impromptu "When you have a radio or victrola in your home, is it worthwhile to play a musical instrument?" (Elliot, 2005, p. 81). The essay assessment allowed students to choose from multiple prompts. These prompts set a relatively modest length of about 350 words and gave students one hour to complete them. Brigham, however, was unhappy with reader reliability, which was extremely low (Brigham, 1934). (Reliability refers to the consistency of measurement. All measurements contain some amount of error, but multiple measurements with high reliabilty have only very small and inconsequential differences among them, while the differences in multiple measurements with low reliability will vary substantially.)

As Huot notes (2002), the whole psychometric community was obsessed with reliability, especially, in the case of writing assessments, at the expense of validity. After World War II, inter-rater reliability was achieved by limiting students to a single essay prompt, scoring the essays on a rubric based holistic scale, and severely limiting the time allowed students to write the prompt (Diederich, 1974; Godshalk, Swineford, & Coffman, 1966)

SHORT TIME FOR WRITING ENABLES LENGTH TO BE MAIN PREDICTOR OF SCORE

The quotation from Attali & Burstein at the beginning of this chapter offers strong evidence that this reliability in grading short timed impromptu writing tests, be it inter-rater reliability or reliability between a machine and a human

rater, is largely a function of length. This evidence is corroborated by the comparison of data from various College Board Research Reports (Breland, Bonner, & Kubota, 1995; Kobrin, Deng, & Shaw, 2007; Mattern, Camara, & Kobrin, 2007), a recent study by Milo Beckman (2010), and data I have collected from both online and timed writing assessments I have given at MIT. These data are displayed in Figure 1. Simply stated, when students are being asked to write an essay on a subject they may not have thought much about in a very short amount of time, length becomes the major determinant of the holistic score. However, the function is negative and exponential. Although length appears to predict 40-60% of the shared variance for essays written in 25 minutes, as the time allotted increases, the correlation between length and score decreases significantly. When students have one hour to write, the shared variance predicted by length decreases to approximately 20%, and when students are given 72 hours, length predicts 10% or less of the shared variance of the holistic score.

These findings are also supported by the review of studies of the effect of length and score by Powers (2005). In particular, the effect of length appears to diminish significantly when students are asked to write about something they know about. A study of untimed essays with a word limit of 1,250 words written for a first-year undergraduate psychology class displayed a shared variance between grade and length of only 1.7% (Norton, 1990). These results reflect both common sense and observations from years of evaluating student papers. Writing tasks, not only in composition classes but also in most academic and professional contexts are given with an explicit range of appropriate length (e.g., 250-300 words; 2000-2500 words; or five to seven pages). Almost all writing falls within the specified range, and more often than not, longer papers within the specified range are, in the aggregate, no better than shorter papers. Indeed, it is a fairly unique feature of the timed impromptu that there is no specified length, reinforcing the sense that the student does better who spews out the most words regardless of content or coherence. Moreover, it is similarly apparent that students writing on a subject they know in advance also reduces the influence of length on score.

AES AND CONSTRUCT VALIDITY

The inescapable fact that there is such a close correlation between length and holistic score has not prompted questioning by those involved in Automated Essay Scoring about the validity of the timed-impromptu as a measure of writing ability. Rather, it has prompted them to argue that Automated Essay Scoring can achieve better construct validity than human readers because human raters

are unreliable and sometimes capricious evaluators (Attali & Burstein, 2006; Ben-Simon & Bennett, 2007). They do, however, admit that construct coverage still needs improvement (Quinlan, Higgins, & Wolff, 2009). This chapter will focus on the construct validity of e-rater 2.0 because the Educational Testing Service has been more transparent than the other developers of Automated Essay Scoring—Vantage Technologies and Pearson Education—in describing the specific features that constitute its scoring algorithm.

Although most of the publications by ETS define e-rater's score as *holistic,* the score is no sense the holistic score defined by White in his seminal article, "Holisticism" (1984). The "holistic" score derived by e-rater is, in reality, a weighted sum of analytic scores and sub-scores that fall into five broad categories: organization, development, lexical complexity, topic specific vocabulary usage, and grammar, usage, mechanics, and style. (Attali & Burstein, 2006; Ben-Simon & Bennett, 2007). Quinan, Higgins, and Wolff (2009) argue that these categories map onto the National Writing Project's 6 + 1 Analytic Writing Continuum that was originally based on the categories of 1) Ideas and Content; 2) Organization; 3) Voice; 4) Word choice; 5) Sentence Fluency; and 6) Conventions, but they offer no evidence to support such a claim. A closer analysis of the metrics used for each of the five e-rater categories highlights the basic limitation of all Automated Essay Scoring. They do not understand meaning, and they are not sentient. They do not react to language; they merely count it.

The organization and development metrics are based on the concept of the "discourse element," which derives from the structure of the traditional five-paragraph essay (Attali & Burstein, 2006). The sole metric for organization is the number of discrete discourse elements in the essay such as "thesis, main ideas, and conclusion" (Ben-Simon & Bennett, 2007, p. 10). Operationally, a discourse element is usually seen as a paragraph, with the introductory and concluding paragraphs having a slightly different structure than the middle supporting paragraphs.

It assumes a writing strategy that includes an introductory paragraph, at least a three-paragraph body with each paragraph in the body consisting of a pair of main point and supporting idea elements, and a concluding paragraph. The organization score measures the difference between this minimum five-paragraph essay and the actual discourse elements found in the essay. Missing elements could include supporting ideas for up to the three expected main points or a missing introduction, conclusion, or main point. On the other hand, identification of main points beyond the minimum three would not contribute to the score (Attali & Burstein, 2006, p. 10).

E-rater is so wedded to the structure of the five-paragraph essay that it intentionally will not recognize more than three "supporting points," which trans-

lates as the traditional three supporting paragraphs. While the feature *organization* is defined as the number of discourse elements, development is defined as average length of each discourse element in words. E-rater, and possibly, other machine scoring algorithms, equates length with development. It is not surprising then, that two ETS researchers, Attali and Powers (2008), found that the correlation between both organization and development and overall number of words was so strong, that they could just substitute length in words for both development and organization.

Yet common sense tells us that development and organization are much more complex features than mere verbiage. A horde of rambling unconnected sentences does not develop an idea. Development is the modern equivalent of *Inventio,* Invention, one of the five departments of Classical Rhetoric. However, AES does not know Aristotle, Cicero, or Quintilian. Again, all the machine can really do is count.

Similarly, the two metrics that constitute e-rater's notion of "lexical complexity" are not complex but entirely mechanical and reductive. The first metric simply judges the complexity of words by counting their letters. The longer the word, the more complex it is, replicating the same bizarre logic that determined that the longer a paragraph is, the more developed it is. The second metric is even more curious. It counts the number of words that are infrequently used in a large representative corpus of English prose. Consequently, e-rater rewards the use of jargon and obscure and pretentious language.

These constructs, however, directly contradict the most widely accepted standards for common English prose, although, of course, different discourse genres diverge on specific features. In most contexts, however, brevity is preferred to verbosity, and simplicity preferred to pretentious diction. As Gowers in Chapter 7 of the *Complete Plain Words* (1954) states, "If the choice is between two words that convey a writer's meaning equally well, one short and familiar and the other long and unusual, of course the short and familiar should be preferred." Similarly, the sixth principle of composition in Strunk and White is "Omit needless words" (Strunk & White, 1962, p. 26). Orwell in "Politics and the English Language" (1945) admonishes the reader to avoid pretentious diction and "never use a long word where a short one will do."

These three authors, of course, represent a notion of single standard style for acceptable writing. Recent work has shown that many of the common rules given by these authors, such as to avoid the passive voice, are in direct conflict with common genres of different discourse communities. Scientists and engineers, for example, often prefer the passive voice because it reinforces their activities as observers of objects. Scientific and engineering genres also prefer jargon particular to the specific genres and discourse communities as a short-

hand for communication with audiences who are familiar with those particular concepts. In most, if not all modern genres of written English, however, brevity is preferred to verbosity and simplicity to polysyllabic words. In business discourse, for example, the one or two page memorandum is norm. Less is more. In addition, there are few, if any genres that would, like e-rater, prefer *plethora* and *myriad* to *many* and *egregious* to *bad*.

The last two vocabulary metrics measure "Prompt-specific Vocabulary Usage." This technique is similar to the "Bag of Words" algorithms used by Latent Semantic Analysis, Probabilistic Latent Semantic Analysis, and Naïve Bayes approaches (Rosé, Roque, Bhembe, & VanLehn, 2003). In essence, the machine goes through each sentence looking for specific vocabulary based on the assumption that similarly scoring essays will contain similar vocabulary. With e-rater, there are two distinct metrics. The first metric evaluates an essay, based on a graded sample set of essays, on which numerical score category contains essays with similar vocabulary. The second metric compares the vocabulary of the essay to those of highest scoring essays in the sample set (Attali & Burstein, 2006). These two features, however, ignore the crucial relationships among words that are crucial to meaning. They, in essence, are looking for certain "buzz" words without regard to whether they make any sense. Many six-point essays written to a specific prompt, for example, may contain the word entrepreneurship. However, training students to use such words without caring that they are using them properly, which is what e-rater does, is not improving students' writing skills; it is teaching them to value and write meaningless verbiage with little consideration of content.

GRAMMAR, USAGE, MECHANICS, AND STYLE

In addition to the features outlined above, e-rater evaluates grammar, usage, mechanics, and style by assessing sets of sub-features such as pronoun errors, sentence fragments, subject verb-agreement, article errors, spelling errors, punctuation errors, too many long sentences, too many short sentences, the repetition of words, and the use of the passive voice (Quinlan, Higgins, & Wolff, 2009). These abilities to identify these types of errors in English prose, of course, are not an innovation of e-rater, but rather, e-rater's grammar checking software is just a recent addition to a collection of software that goes back to Writer's Workbench, Grammatik, Correct Grammar, and Right Writer. In the early 1990s, the two leading word processing software packages, Microsoft Word and Word Perfect incorporated highly sophisticated grammar and style checking software that not only identifies problems in spelling, grammar, and

style, but allows users to the option of having the system automatically correct obvious and unambiguous spelling errors. In addition, from 1995 onwards, MS Word not only offers possible corrections for some errors along with offering the user an explanation of the grammatical or stylistic rule.

Microsoft's Grammar Checker (MSGC) was developed and is maintained and improved by the Natural Language Processing Group of Microsoft Research, which consist of approximately fifty computational linguists. But although much more sophisticated than earlier grammar checkers and backed with enormous resources for continuing development, the MSGC is still often capable of giving very bad advice. The anomaly noted in Word 2000 by McGee and Ericsson (2002) still exists in MS Word 2007. If I write that Bill was left by the side of the road, MSGC still suggests to change it to "The side of the road left Bill." Recently, Herrington and Moran (2012), have demonstrated significant flaws in e-rater and Criterion. The system marks perfectly correct parts of sentences as grammatical errors.

This digression on Microsoft Word's grammar and style checker is meant to demonstrate that the grammar and style algorithms in specialized programs such as e-rater will never have the sophistication and continuing improvement of MSGC, which still possesses substantial limitations. The reason is simply a matter of scale. Millions of copies of MS Word are sold every year, more than enough to support a large team of computational linguists constantly improving the product. The combined customer base of all three major AES systems, Intellimetric, Intelligent Essay Assessor, and the Educational Testing Service's e-rater is a miniscule fraction by comparison.

CONCLUSION

There are, then, four interrelated points, that argue strongly against the use of AES both as an assessment tool and as an aid in instruction. First, the "holistic" score produced by AES is largely a function of the length of the essay. Second, the abnormal nature of the short timed impromptu writing test produces this strong correlation of length to score. This strong correlation does not appear in prose in which the student either knows the subject beforehand or has had sufficient time to write. Third, the metrics employed by programs like e-rater do not reflect the constructs they are supposed to measure. They are largely irrelevant at best, and sometimes counter-productive at worst. Finally, the grammar checking and instructional function of e-rater and Criterion are much more limited than the much more developed functions in standard software such as MS Word, which itself has major limitations.

E-rater, and probably the two other major AES engines Vantage Technologies' Intellimetric®, and Pearson's Intelligent Essay Assessor primarily perform two actions: they imperfectly count errors and count words and characters with unerring precision. This counting is the real construct informing AES. Often the underlying, but unstated, motive in assigning timed impromptu writing test is to elicit errors from students and count them. A low density of error, that is, the longer the student text and the fewer errors in it quickly becomes the unstated but very real construct that underlies this kind of assessment. Yet the past thirty years of writing studies, beginning with Mina Shaughnessy (1979) reveal that command of grammar, mechanics, topic specific vocabulary, and sentence complexity are an integral part of a complex set of socio-cognitive processes.

For AES to be valid, it must incorporate valid constructs and accurate measures of those constructs. Developers of AES systems say that these constructs must come from writing teachers (Attali & Burstein, 2006; Ben-Simon & Bennett, 2007; Quinlan, Higgins, & Wolff, 2009). Yet AES systems measure a construct that bears no relation to the well-articulated abilities enumerated in the recent *Framework for Success in Postsecondary Writing jointly* developed by the Council of Writing Program Administrators, the National Council of Teachers of English, and National Writing Project (2011). This *Framework* clearly articulates the construct that needs be measured to assess writing ability: the rhetorical ability to integrate an understanding of audience, context, and purpose when both writing and reading texts; the ability to think and obtain information critically; the ability to effectively employ multiple writing strategies; the ability to learn and use the conventions appropriate to a specific genre of writing; and the ability to write in various and evolving media. There is no construct of AES that comes close to assessing these skills.

Portfolio evaluations clearly offer the most promising platform for assessing this complex construct. But there are other more limited platforms that, at least, come much closer than AES, and as technology advances there will be others. The iMOAT system and similar online systems, for example allow for a much greater construct validity in that they assess students' engagement with texts, their ability to think critically for more than five minutes, and their ability engage in all stages of the writing process (Peckham, 2006; Peckham, 2009; Peckham, 2010; Perelman, 2004). Other, more advanced platforms will evolve. It is almost certain, however, that the prose written on these platforms will not be amenable to grading by machine until several significant revolutions occur in both theoretical and applied linguistics, until there is a theoretical framework for semantics that will allow a computational implementation, until machines understand meaning. Until then, all AES will be is reductive counting.

NOTE

1. I want to thank Norbert Elliot, Suzanne Lane, Charles Bazerman, and the anonymous reviewers who helped me immensely in focusing this chapter and providing me helpful and crucial suggestions.

REFERENCES

Attali, Y. (2007). *Construct validity of e-rater in scoring TOEFL essays* (ETS Research Rep. No RR-07-21). Princeton, NJ: Educational Testing Service.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *The Journal of Technology, Learning, and Assessment , 4*(3). Retrieved from http://escholarship.bc.edu/jtla/vol4/3/

Attali, Y., & Powers, D. (2008). *A developmental writing scale* (ETS Research Report RR-08-19). Princeton, NJ: Educational Testing Service.

Beckman, M. (2010). Quality vs. quantity: How to score higher on the SAT essay component (Unpublished manuscript).

Ben-Simon, A., & Bennett, R. E. (2007). Toward more substantively meaningful automated essay scoring. *The Journal of Technology, Learning and Assessment , 6*(1), Retrieved from http://escholarship.bc.edu/jtla/vol6/1/

Biola, H. (1982). Time limits and topic assignments for essay tests. *Research in the Teaching of English, 16,* 97-98.

Breland, H., Bonner, M., & Kubota, M. (1995). *Factors in performance on brief, impromptu essay examinations* (Report 95-4). New York: College Board.

Brigham, C. C. (1934). *The reading of the comprehensive examination in English: An analyis of the procedures followed during the five reading periods from 1929-1933.* Princeton, NJ: Princeton University.

Chodorow, M., & Burstein, J. (2004). *Beyond essay length: Evaluating e-rater's performance on TOEFL essays* (Research Report 73). Princeton, NJ: Educational Testing Service.

Council of Writing Program Administrators, National Council of Teachers of English, and National Writing Project. (2011). *Framework for Sucess in Postsecondary Writing.* Retrieved from http://wpacouncil.org/files/framework-for-success-postsecondary-writing.pdf

Diederich, P. B. (1974). Measuring Growth in English. Urbana, IL: NCTE.

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

Godshalk, F. I., Swineford, F., & Coffman, W. E. (1966). *The measurement of writing ability*. New York: College Entrance Examination Board.

Gowers, E. (1954). *The complete plain words.* Retrieved from http://www.ourcivilisation.com/smartboard/shop/gowerse/complete/index.htm

Herrington, A., & Moran, C. (2012). Writing to a machine is not writing at all. In N. Elliot & L. Perelman (Eds.), *Writing assessment in the 21st century: Essays in honor of Edward M. White.* New York: Hampton Press.

Huot, B. (2002). *(Re)articulating writing assessment*. Logan, UT: Utah State University Press.

Kobrin, J. L., Deng, H., & Shaw, E. (2007). Does quantity equal quality?: The relationship between length of response and scores on the SAT essay. *Journal of Applied Testing Technology, 8*(1)*,* 1-15.

Mattern, K., Camara, W., & Kobrin, J. (2007). *SAT® writing: An overview of research and psychometrics to date*. New York: College Board.

McGTee, T., & Ericsson, P. (2002). The politics of the program: MS Word as the invisible grammarian. *Computers and Composition*, *19,* 453-70.

Norton, L. S. (1990). Essay-writing: What really counts? *Higher Education*, 20, 411-442.

Orwell, G. (1945). P*olitics and the English language.* Retrieved from http://mla.stanford.edu/Politics_&_English_language.pdf

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education, 61*(4)*, 127-142.*

Peckham, I. (2010). Online challenge versus offline ACT. *College Composition and Communication 61*(4), 718-745.

Peckham, I. (2009). Online Placement in First-Year Writing. *College Composition and Communication 60*(3), 517-540.

Peckham, I. (2006). Turning placement into practice. *WPA: Writing Program Administration, 29*(3) 65-83.

Perelman, L. (2004). *Assessment in cyberspace.* Retrieved from http://www.mhhe.com/socscience/english/tc/perelman/perelman_module.html

Powers, D. (2005). *"Wordiness": A selective review of its influence and suggestions for investigating its relevance in tests requiring extended written responses.* Princeton, NJ: Educational Testing Service.

Powers, D., & Fowles, M. (1997). *Effects of applying different time limits to a proposed GRE writing test* (Research Report 96-28). Princeton, NJ: Educational Testing Service.

Powers, D., Burstein, J., M., C., Fowles, M., & Kukich, K. (2001). *Stumping e-rater: Challenging the validity of automated essay scoring* (Research Report 01-03). Princeton, NJ: Educational Testing Service.

Quinlan, T., Higgins, D., & Wolff, S. (2009). *Evaluating the construct-coverage of the e-rater scoring engine*. Princeton, NJ: Educational Testing Service.

Rosé, C. P., Roque, A., Bhembe, D., & VanLehn, K. (2003). *A Hybrid Approach to Content Analysis for Automatic Essay Grading.* Retrieved from http://acl.ldc.upenn.edu/N/N03/N03-2030.pdf

Shaughnessy, M. P. (1979). *Errors and expectations*. New York: Oxford University.

Strunk, W., & White, E. B. (1962). *The elements of style* (2nd ed.). New York: Macmillan.

White, E. M. (1995). An apologia for the timed impromptu essay test. *College Composition and Communication , 46*(1), 30-45.

White, E. M. (1984). Holisticism. *College Composition and Communication, 35*(4), 400-409.