2015

Automated Essay Evaluation and the Computational Paradigm: Machine Scoring Enters the Classroom

Catherine M. Barrett
University of Rhode Island, kenna_b@yahoo.com

AUTOMATED ESSAY EVALUATION AND THE COMPUTATIONAL PARADIGM: MACHINE SCORING ENTERS THE CLASSROOM

BY

CATHERINE M. BARRETT

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

RHETORIC & WRITING

UNIVERSITY OF RHODE ISLAND

2015

DOCTOR OF PHILOSOPHY DISSERTATION

OF

Catherine M. Barrett

APPROVED:

Dissertation Committee:

Major Professor Robert Schwegler

Julie Coiro

Renee Hobbs

Abrán Salazar

Nasser H. Zawia
DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND

2015

ABSTRACT

Grading papers is a "tedious, repetitive, and time-consuming" undertaking, one that invites the sort of sympathies one might receive upon the "death of a pet" (Baker, 2014, p. 36). Perhaps, though, the only thing more distasteful for an English professor than having to grade hundreds of essays each semester is having to hand the job over to the likes of a mindless software program. Automated Essay Evaluation (AEE), the process of scoring essays by computer, was developed in the 1960s but has mostly aroused suspicion and derogation from the composition community.

The parable of the blindfolded villagers describing the same elephant from different parts of its body serves as an apt metaphor for how rhetorical choices screen in or screen out different facts about AEE. Each research camp reports on a different facet of AEE, but it is possible for these multiple realities to describe the same beast.

This dissertation seeks to: (1) describe the rhetorical contours of arguments for and against AEE, both currently and historically, exposing their limitations and motivations; (2) explore through a small study how differential frames of data analysis affect interpretations of AEE's utility, namely how error analysis adds an emphasis on individual students rather than student-aggregates; and (3) sketch ways that attention to errors is instructive in understanding the pitfalls of both human and machine scoring of essays. Because human scoring—despite its drawbacks—is still the superior method, I end by suggesting that AEE would be best used when the economic choice is between machine scoring and feedback, or no scoring and feedback at all. The context of online learning in globally disadvantaged populations is one such example.

ACKNOWLEDGEMENTS

Above all, thanks go to Bob Schwegler, whose encouragement fostered my own determination. I'm grateful to members of my core committee Julie Coiro, Renee Hobbs, Mike Pennell, and especially to Abrán Salazar for his review of the quantitative aspects of the project. Les Perelman helped shape my research plan and combed through my data with me on several afternoons at the Dartmouth Summer Seminar in Composition Research. The members and advisors of the Dartmouth Summer Seminar itself were terrific, particularly Chris Anson, Chuck Bazerman, and Tiane Donahue. My URI colleagues Tim Amidon, Jeremiah Dyehouse, and Nedra Reynolds have helped in various ways as did other colleagues who attended my Turing presentation in November 2013. Also deserving thanks are my URI doctoral cohort members Marcy Isabella and Jay Peters, oral defense committee members Kim Hensley-Owens and Sandra Ketrow, who provided useful commentary, and participating raters Allison Bass, Sarah DeCapua, Adrienne Dowd, Lois Lake-Church and Wendy Grosskopf. Stephanie Fischer and Matt MacKnight also participated as instructors. I enjoyed early-stage conversations with Andrew Klobucar. At ETS, Chinni Ramenini and Susan Yetman greatly helped facilitate my research. Through their IRBs, Southern Connecticut State University and the University of Rhode Island made it possible to conduct my assessment study. Donna Hayden and Michelle Caraccia demystified the whole doctoral process.

Without tremendous support from home, my dream to return to school after 13 years in the private sector would never have been possible. Thanks to Vikram and Cathie, Savannah and Emma L. Others who wrote with me, counseled me, and cheered me on include Amber Batata, Dennis Barrett, Jade Barrett, Mike Barrett, Resha Cardone, Mira

Debs, Anna Mubarak, and Claire Rutledge. My thanks to all the above and to those whom I may have inadvertently left out.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Writing assessment used to measure writing quality, writing ability, placement, course outcomes, and even to grade and comment on student papers is that particular focus of researchers and scholars in the writing assessment community. Drawn from a variety of disciplines, including writing studies, English education, and psychometrics, the members of this community have much of importance to say and do in the current school and college environment, characterized by frequent high-stakes testing and attention to placement course outcomes. The writing assessment community is small but powerful as a result of the quality of its research and its influence over nationwide testing policies and programs.

Perhaps the most potent disagreement dividing the writing assessment community at present is over software programs that score student essays, a practice known as automated essay evaluation (AEE). At one pole, AEE supporters claim that computers can rate papers accurately, helpfully, and efficiently. At the opposite pole, proponents of human raters argue that the machines are relatively useless and that human raters alone should be used.

This study begins and ends with a position in between. It critiques various ways of framing the AEE debate, because AEE has become the elephant seen from many perspectives. Both overly strong arguments for AEE and their hasty dismissals present one-sided, rather than multifaceted cases for their claims. For instance, chapter 2 investigates the way in which AEE has been framed by its founder Ellis Page as a version of a Turing test, a move to secure legitimacy by appealing to Alan Turing's *ethos*. That chapter then looks at the way in which opponents of AEE have framed the practice as

impersonal, mechanical, and dehumanizing, a way of undermining legitimacy via the *ad hominem* (really an *ad machinam*) attack. Then, in reviewing literature on AEE, chapter 3 casts the current day debate as one where proponents highlight reliability of machine scoring while opponents foreground validity. Far too often, the rhetorical positioning of various camps in the field obscures the fact that there are as many reasons to be open to AEE as there are to resist it. Chapter 4 takes a small empirical study and pushes the framing of the data analysis beyond the typical agreement measures often provided, to measures of distribution, correlations of score and essay length, and a version of error analysis that highlights failures in agreement as much as successes.

Key Assumptions of Computer Scoring

Underlying the claims for AEE, which are now dominant in certain areas of the field, is a series of complicated assumptions. These assumptions occupy three spaces in what is called the "communication triangle" of writer, reader, and text. One of the first assumptions underlying AEE involves the reader: that there are a set of traits in a text that teachers value and use to judge the quality of student writing. Diederich, French, and Carlton (1961) at the Educational Testing Service (ETS) came up with five factors through a study of 300 essays that they concluded mattered to teachers assessing writing. Although this study concluded that over half of variation in scores was due to error/random variation, trait analytical grading became dominant as a result of this study, solidifying the narrative that readers make scoring decisions based on textual analysis and not extraneous factors.

The second assumption, involving the texts themselves, that has made AEE possible is that these features of writing that were said to matter to teachers could be considered "intrinsic" qualities that could be approximated by a computer program representing these features through proxy measures or "proxes." Slotnick (1972) identified a number of computer proxes that were characteristic of six intrinsic factors in judging writing quality. The thought was that if the proxes lined up sufficiently with the intrinsic qualities, the computer could not only reliably assess writing, but do so validly as well.

Building on these founding assumptions that there exist separable features that teachers use to judge essay quality, and these features can be captured by software, Page (1966) was able to demonstrate a correspondence in outcome between raters who followed this rubric in rating papers and machines that followed a similar rubric. Although Page was unable to achieve an interrater reliability much higher than .5, recent AEE proponents have demonstrated to their satisfaction that they can achieve an interrater reliability of .7 or .8, and in so doing, can equal the reliability of human raters.

The third assumption of AEE involves writers: that these texts produced by students under test conditions are actually representative of a stable and universal construct of student writing ability across multiple genres. That is to say, scores along these generally short, timed essays are said to be meaningful indicators of a student's writing ability.

This tripartite storyline has been challenged by those who oppose AEE. That teachers' judgments are in fact based on intrinsic analytic traits has been challenged by Broad (2003) and others. That proxes and trins are closely enough related so as to largely

reduce validity to reliability has been challenged by Condon (2013). That there is anything like a general writing ability tapped into by short timed impromptus rated by computers has also been taken up by critics (e.g., Perelman, 2012b). This has become a construct validity problem and it appears to be a serious one for AEE. However, AEE proponents can concede that impromptu tests may not measure a full writing construct while still maintaining that they have a place and time.

Presenting empirical data comparing human and machine score performance has become the standard mode of responding to critiques. Researchers present measures such as human and machine test set means, test set standard deviations, percent agreement between raters, Pearson correlation of scores, and Cohen's kappa coefficient. The point they seek to make is that no matter what the psychology of human scoring, and no matter what the computational basis of machine scoring, the scores match each other to a close enough extent. The process just seems to work. Questions of mechanism have, in the parlance of Latour (1999), become black-boxed. Automated approaches have continued on seemingly unruffled by the theoretical controversies.

Indeed, AEE approaches continued to gain traction over the latter part of the 20th and first decade of the 21st centuries. At the time of this writing, scoring engines are used in the essay sections of standardized examinations including the TOEFL, the GMAT, and the GRE. Since many standardized tests are scored by two readers, automated engines often serve as the "second reader." Based on publicly-available figures of 800,000 annual takers of the TOEFL exam, 700,000 GRE takers, and nearly 300,000 GMAT takers, it would seem that on the order of 1.8 million students a year come into contact with computer scoring, through these three examinations alone. With the PARCC and

SMARTER BALANCED consortia—the two major administrative bodies for the new Common Core curriculum—declaring their interest in automated scoring, millions of Kindergarten through twelfth-grade students also stand to come into contact with automated scoring.

And, AEE has made advances in higher education. "Machine scoring," wrote Patricia Ericsson and Richard Haswell, "no longer has a foot in the door of higher education. It's sitting comfortably in the parlor" (2006, p. 4). In the college setting, AEE applications function as placement technologies in first-year composition as well as feedback-providing platforms. According to the College Board's website, 1,500 secondary and post-secondary institutions use ACCUPLACER, that company's writing course placement application, covering 2.5 million students each year. Public reports indicate that at least the following institutions have adopted the Educational Testing Service's online writing tool, *Criterion*® (which offers automated essay assessment and feedback) in the past five years alone: Fresno State University (2010), New Jersey Institute of Technology (2010), Jackson State University (2012), San Jose State University (2012), and the University of North Carolina-Wilmington's Business School (2008). According to its website, Wilmington University (DE) uses *Criterion* as an admissions tool to the Doctor of Education program. Regent University (VA) reports a similar use of *Criterion* for admission into its Master's programs.

The Collegiate Learning Assessment (CLA+) is scored by a combination of Pearson's Intelligent Essay Assessor (IEA) and human readers; each essay receives an automated and a human score. The earlier version of the CLA+, known as the CLA, is reported to be in use at over 700 institutions of higher education.

While proponents have relied on large-scare quantitative studies and detractors on qualitative evidence, what has been little attempted is a study that presents opportunities for both. It is exactly because AEE performs best when scores are aggregated and appears in the worst light when considering how it treats individual cases that these separate research paradigms have evolved. This project takes a "both-and" approach.

In addition to studying statistical analyses indicating how the computer performs on average compared to human raters, I ask how a computer would score individual essays, since it is the individual student and her essay that is of importance to writing studies. My study points out that even under statistically solid testing conditions, there will be students who stand to have their essays wrongly scored by the automated system. To take a simple example, even if the exact agreement between a human and machine rater is .9, a very high ratio, ten of 100 students may have essays that have been mis-scored. I say "may have" because without a second human rater, there is no way to know whether the human score or the machine score misses the mark. The fact is that in any system of approximation, there will be error. The question is how much error is acceptable.

In building up to my case that not only statistical averages but individual cases matter for the AEE dialogue, and my attempt to sketch the implications of this view, I begin with a look at the early statistical arguments put forward by Ellis Page, the developer of the first AEE system. I use computational theory to explore the way in which the proposed commonalities between machine and human rating are based in aspirational metaphors as much as unequivocal evidence.

Foundational Moments of AEE

In 1966, education researcher Ellis Page issued a bold proclamation. In an article entitled "The Imminence of Grading Essays by Computer," he described a new computer program that was to save English teachers the tedium of grading student essays by hand. Page augured that "[W]e *will* soon be grading essays by computer, and this development *will* have astonishing impact on the educational world" (p. 238, emphasis his).

Like mine, Page's case for computer scoring of essays began with students—specifically, secondary school students. While everyone should of course be assigned daily or weekly writing themes that receive thoughtful commentary from well-trained English teachers, what they often get is hasty comments from overburdened staff, on papers written at much less frequent intervals. The less-than halcyon situation in college English courses reproduces the traffic jam of the lower years. "The system is not fair to [English teachers]," Page concluded. "It rewards them no more than others, while requiring them to accept a staggering responsibility with neither the time nor equipment to meet it…English teachers are (after students) the most obvious victims" of this system (1966, p. 259). Computer scoring was, for Page, the solution to the time constraints that teachers faced during grading, given a system unwilling or unable to invest more resources into the teaching mission.

Page's discussion raises a concern that has only intensified since his writing: a concern that the "system"—roughly, the set of institutions in society and the laws that regulate them—is unfair to students and teachers of English in its denying of teachers the material resources necessary to do the job necessary for students to develop sufficient competence in writing. Ongoing systemic problems are certainly at the forefront of the

writing studies profession as a cursory glance through the National Council of Teachers of English (NCTE) position statement library reveals titles such as "Why Class Size Matters Today," "Statement on the Status and Working Conditions of Contingent Faculty," and "Resolution on Challenging Current Education Policy and Affirming Literacy Educators' Expertise."

If these were the motives for developing computer programs that graded essays—improved learning outcomes, better working conditions and greater parity for writing teachers, and increasingly valid forms of assessment—they brook no controversy. But controversy was what Page expected and conceded that "for many of us, the idea at first seems utter nonsense," in the opening paragraph of his (1966) essay. He went on to suggest that computerized scoring had additional applications in the field of psychometrics—educational measurement—because it would, he asserted provide a way of reliably and validly measuring essay quality.

Controversy was what Page did in fact get. By 2013, the NCTE itself had issued a position statement that said that "Yet when we consider what is lost because of machine scoring, the presumed savings turn into significant new costs -- to students, to our educational institutions, and to society" (para. 5). Implied in NCTE's position statement is the assumption that students and faculty would *not* stand to benefit from the applications of this technology.

There is a paradoxical nature to this particular controversy in which the stated motivations for and against computer scoring of essays appear largely the same: helping students, tied to suspicions of an unjust system or global order that ultimately damages

vulnerable students and their also vulnerable teachers. The reasoning *for* automated evaluation is strikingly similar to the arguments *against* it.

How did something proposed to benefit overworked teachers and their students years ago morph into a threat to these very actors? What, in fact, *are* the benefits or the costs of automated scoring on students? In the intervening years since Page's germinal essay, automated scoring has been closely studied. Many have considered whether students stand to benefit or not from this technology. My project focuses on a particular operationalization of "benefit": whether the scoring can be considered to be accurate in college settings, and what the implications are of scoring "errors." I target my work toward the writing assessment community as another perspective for scholars and faculty called to advise their institutions of the benefits or harms of automated assessment. By calling attention to framing, this projects situates itself not as "the" response to AEE but as one of many perspectives adding up to a fuller conception.

Before continuing, a few reflections on my own subject position. As a graduate student in a field where less than half the new PhDs get tenure-track positions, anything threatening to further disrupt the fragile labor market could be perceived as undesirable. Whether automated scoring has actually or would result in job elimination is unknown. But teachers understandably cast wary eyes; administrators do not need much pretext to cut teaching sections or jobs. Is my position conflicted? If so, it is not the only one, given that many of the studies on computer scoring have been undertaken by software developers.

In addition to my professional identity is my scholarly orientation: I was trained at the undergraduate level as a cognitive scientist, at the graduate level as a philosopher of

science, and then finally at the doctoral level as a compositionist. I am intellectually curious about automation, its limits, and its promise, distinct from any subliminal fears of its ascendance. I have chosen to frame the empirical piece of my study around the impacts of AEE for students because they are still, as they were for Page, the original victims of an educational regime insufficiently supportive of teachers and the human capital required to truly teach writing.

Given my scholarly orientation, this study does not offer a Marxian or cultural critique of technology or of AEE; it does not offer a Foucaultian genealogy locating it in the history of the "examination"; nor does it otherwise focus much attention on the implications of AEE for the valid concerns that the NCTE has raised about societal impact. It may certainly be that our education industry has created and perpetuated automated essay evaluation for commercial gain—but doing justice to those questions would require economic data and more. Setting them aside should not be seen as an implied devaluing; indeed they call for serious and methodical inquiry beyond the scope of this paper.

Where did automated scoring end up? By 2003 education researchers Gregory Cizek and Bethany Page felt comfortable enough with the state of the technology to respond to Page's prophesy by writing that "automated scoring of extended responses to test items or prompts is a *fait accompli*" (p. 118). From an imminence to a *fait accompli* in 37 years: together, these quotations could be said to project a rhetoric of technological destiny for automated scoring, a master narrative framing technological innovation as inexorable, virtuous, and "profitable to mankind as a whole" as Lyotard (1986/2010) has said in describing the modernist project (p. 1467).

As to the origins of such technological progress, some might say that these technologies are, as Donna Haraway (1985/2010) pointedly wrote of the cyborg, "the illegitimate offspring of militarism and patriarchal capitalism, not to mention state socialism" (p. 2192). Haraway's metaphor of the cyborg (a literal hybrid of human and machine) describes the fusing of the technological and the biological, and her ultimately balanced stance toward cyborg reality offers a lens for envisioning the unavoidable technology of automated essay evaluation.

In a sharp response to the flowering of automated approaches, a group of college composition professors (including many members of the Conference on College Composition and Communication) circulated an online petition signed by 4,108 people as of March 2013, including K-12 teachers, college teachers, community representatives, and public intellectuals including linguist Noam Chomsky. The *humanreaders.org* petition called automated scoring "a major disservice to all involved," going on to term automated scoring as "trivial," "reductive," "inaccurate," "undiagnostic," "unfair," and "secretive, with testing companies blocking independent research into their products" (2013, para. 2). Here again we see a reference to corporate interests, underlining the suspicion articulated by Haraway that AEE is a creation of capitalism and the state apparatus that sustains it.

This dissertation takes a degree of refuge in Haraway's cyborg. For automated essay evaluation is itself a cyborg: as some AEE opponents may not have considered closely, its learning algorithms issue score predictions gleaned from training on thousands of similar human-scored essays. The human scores serve as the basis for the machine predictions. Machine scores are, thus, scores made in our own image, or at least

approximations of that image. Recognizing the complicities of the cyborg, Haraway seeks not necessarily to resolve dialectical thinking about technology, but rather to contend that it is politically necessary to hold multiple and contradictory viewpoints regarding posthumanism. After all, as she notes, "illegitimate offspring are often exceedingly unfaithful to their origins" (p. 2192). We must disentangle the genealogy of these technologies from their ultimate manifestation in society.

In her exegesis, Haraway marks out two perspectives that the posthuman world could potentially be about, the first being "the final imposition of a grid of control on the planet, about the final abstraction embodied in a Star War apocalypse waged in the name of defense, about the final appropriations of women's bodies in a masculinist orgy of war" (1985/2010, p. 2196). This might be akin to the suspicious postmodern perspective of Lyotard.

But the second perspective, Haraway goes on to say, is that "a cyborg world might be about lived social and bodily realities in which people are not afraid of their joint kinship with animals and machines, not afraid of permanently partial identities and contradictory standpoints" (1985/2010, p. 2196). Haraway concludes that "the political struggle is to see from both perspectives at once" (p. 2196). I find the "both…at once" perspective useful because, as Haraway says, "single vision produces worse illusions that double vision or many-headed monsters" (p. 2196). What I take from this reflection is that staking out a position as an "advocate" or an "opponent" filters reality and constructs a rhetorical stance appropriate for that position—a Burkean terministic screen (1966). The filtered reality of the "advocate" position focuses, for instance, on the great reliability that AEE is said to have with human scorers; the filtered reality of the

"opponent" emphasizes the paucity of the writing prompts that could actually produce reliable scores. Both positions can actually be true at the same time—but this is an insight most easily arrived at by someone prepared to stand in both places.

In exploring the question of automating the judgment of writing quality—in standing in both places—I find it useful to employ multiple disciplinary perspectives. I draw upon philosophy of science, rhetoric of science, educational assessment, composition studies, and to a lesser extent critical theory and ethics in undertaking this mixed-method study of AEE. I seek not to reduce the practices of humans writing to machines/machines writing to humans to one Harawayan binary perspective or another. It may certainly be useful to see the propagation of these practices as driven by the capitalist march of testing companies and the socialist insistence on a certain type of educational outcome. However it may be that the illegitimacy of cyborgs enables the technology to stretch past that *telos*. Instead, the ambit of my project is to see "both perspectives at once"—to frame a technology such as AEE first as the result of scientific discoveries that while ideologically bound are themselves worthy of further, and dispassionate, investigation, while simultaneously acknowledging that automation is both produced by and productive of consumer capitalism. In this monograph, I hope to cast automation as a phenomenon with *no* master narrative (or perhaps more accurately, with a multiplicity of such narratives) showing, for instance, how technologists worked alongside teachers to create the case for automated assessment even as physicists found themselves expounding quite similar prose to humanists in attempting to find something unique about mathematical reasoning that would position it beyond the full reach of machines.

As far as what is to follow: the next chapter (chapter 2) is an origins story. It locates the intellectual predecessors of automated essay evaluation in computability and computationalism—theories that arose with the origins of computer science. It is a story upon which Ellis Page draws to undergird his methodology for arguing the merits of computer scoring, and it sets up a method of evaluating computer scoring that has persisted until this day, its ubiquity only now beginning to be questioned. Page's original argument, which has embedded itself into the entire AEE edifice, was that machine scoring was sufficiently human. He felt that in order to legitimate automated scoring, he had to give it a cyborg identity. This "cyborg" argument—the postulation that brains are in some sense computers or conversely, that computers are capable of intelligence—has traces through the way current AEE is represented as comparable to human grading practices. Drawing on the symbolic capital of Alan Turing, Page's focus on mimicry has foreclosed other options for criteria that count as evidence of the "success" of automated scoring. This chapter is part rhetorical study and part philosophy of science. I have chosen to blend these genres while accepting the risk that the normative philosophical piece will undercut the descriptive rhetorical component.

Chapter 3 presents a literature review of current assessment theory and research regarding automated scoring. One conclusion drawn in chapter 3 is that proponents and opponents are both right in a sense: AEE does score reliably in certain contexts; those contexts can be challenged for lacking validity. The framing is key to understanding why the two groups talk beyond each other.

Following the literature review, chapter 4 presents the results of a pilot assessment study comparing the performance of the *Criterion* scoring engine and online

essay evaluation service (called *e-rater®*) to trained human raters using a different rubric more specific to collegiate writing than those used typically. With this study I endeavor to illustrate how Replicable, Aggregable, and Data-supported (RAD) studies (Haswell, 2003) on automated scoring may be designed that are possible to be carried out on smaller scales than the large-scale studies promulgated by industry researchers and institutions that have already adopted automated technologies. The study pilots a rubric derived from the Framework for Success in Postsecondary Writing, (O'Neill, Adler-Kassner, Fleischer, & Hall, 2012).

In sketching directions for future research, chapter 5 is a thought experiment about the nature of error: the ways in which people and machines are bound to err in distinct patterns. The unavoidable nature of error in natural language processing raises suspicions for scoring in general. The nature of machine error is cause for derision, but the relentless consistency of machine scoring opens a window to human fallibility. A cyborg future is one in which both parties must play to their strengths.

As a segue into chapter 2, it would be useful to briefly orient the reader to how automated essay scoring may be defined and how the software typically operates. While precise definitions tend to escape their bounds, a heuristic definition will be offered here in order to motivate the forthcoming discussion. Further chapters will discuss AEE in much greater depth in terms of its underlying technologies and the research on its empirical characteristics. For now the following stipulative definition of AEE will be

used: *automated essay evaluation represents the use of computer programs to evaluate writing*.[1]

Sometimes, the technology is referred to as "automated essay scoring" or "automated writing evaluation," the first to emphasize the summative aspect of the practice, and the second to emphasize its broader applications. While I am sensitive to the fine-grained distinctions inherent in these varied wordings, I have chosen to use "automated writing evaluation," following the terminology used in the most recent industry-produced volume on automated essay techniques, the *Handbook of Automated Essay Evaluation* (Shermis & Burstein, 2013). While names are not neutral signifiers, adopting the sanctioned nomenclature has the advantage of both inviting interrogation of that name as a rhetorical object while positioning me as a researcher working on a common theoretical construct.

Essay evaluation engines are currently employed commercially for three major purposes: high-stakes test scoring, course placement, and providing real-time diagnostic and feedback advice to student writers (for further discussion, see chapter 3). While no scoring engine is the same, and while the architecture of scoring engines will be explored further in chapter 3, a general understanding will be useful here. What happens in most cases is that the systems are programmed to recognize basic features of text, such as whether a word is a noun, verb, or adjective (etc.), the beginning and ending of

[1] This definition essentially follows Shermis, Burstein, and Bursky (2013): AEE is "the process of evaluating and scoring written prose via computer programs" (quoting Shermis & Burstein, 2003). Note that McAllister and White (2006) prefer the term "computer-assisted writing assessment," to "automated essay evaluation," maintaining that computer assessment is not fully *automated* because it depends on humans interacting with machines. However, automation is a key metaphor in the cyborg discourse, even automation that occurs in the kinds of mediated interactions McAllister and White refer to. "Automation" as I use it in the cyborg conversation departs slightly from its original sense of fully *automatic* and lies closer to the idea of an *automaton*, a lifeless entity imbued with human-like qualities.

paragraphs, an essay's overall word count, and so forth. From this raw data the computer can make judgments as to whether a sentence contains a grammar or spelling error, its lexical complexity, how developed its paragraphs are, and so forth. Taking a set of pre-scored essays, the software compares them along all the textual features it can measure, and creates a mathematical score prediction for each one. The score is based on a mathematical formula that works out how the features contribute to the essay's human score. Comparing the predicted score to the pre-assigned score, the learning algorithm can train up, meaning adjust the relative importance of each feature in its formula, and try again to predict the score. Once the computer matches the human scores to a specified degree, it can apply its feature weightings to *new* essays, thereby "scoring" these. This framework of computer scoring agreeing with human scoring on average—the *interrater reliability* of the machines—has become the standard measure of effectiveness of computer scoring. In abstract terms, reliability is a suitable measure for this purpose, but many have charged that it comes at the expense of validity, as we will see in coming chapters. The next chapter traces the early attempts to establish reliability as the key measure for AEE.

In the Educational Testing Service (ETS) engine called *e-rater,* scores are calculated as a "weighted average of the standardized feature values, followed by applying a linear transformation to achieve a desired scale" (Attali, 2007, p. 4). The algorithm is one of multiple regression manipulating predictor variables to best fit the training set.[2] After textual features have been assigned weights based on estimated importance and then placed along the same scale, the algorithm sums the relative

[2] Multiple regression is a mathematical technique that determines the relationship between several predictor variables (in this case, essay features) and an output variable (e.g., a score).

contributions and converts the resulting number to the best-fitting value on the assessment scale, often a 1-6 interval scale.

Such textual features were originally classified into hypothetical “trins” (intrinsic variables of interest) and “proxes” (approximations, or correlates of intrinsic variables). A “trin” might be the aptness of word choice for a given situation, and a “prox” might be the proportion of uncommon words used. Proxes were thought to be those variables that correlated with trins but could be measured by computer (Page, 1966, p. 213). These notions will figure into subsequent chapters.

2. THE COMPUTATIONAL PARADIGM AS A GROUND FOR ESSAY EVALUATION: A METATHEORY

Key developers of AEE, situated against the backdrop of revolutions in computer science, sought legitimacy through certain ways of speaking and writing about automation. These ways of speaking and writing involved likening people to machines (by speaking about humans as if they were computational devices) and likening machines to humans (by speaking about computers as if they were sentient). These rhetorical moves intended to secure legitimacy also extended to tending to emphasize the statistical performance of AEE over its performance in individual cases. Because individual students and individual essays matter greatly in the field of writing studies, the backgrounding of individual score assignments and their implications, and the more general comparing of people to computers, tended to backfire in AEE's quest for legitimacy.

Echoing the position of many compositionists about AEE, Patricia Ericsson (2006) wrote that "if composition is about making meaning—for both the writer and the reader—then scoring machines are deadly" (p. 37). Machines embody an information theoretic approach to meaning, she further argued, concerning themselves with reduction and reproduction of messages rather than with human-like, let alone human, understanding.

In building her case against both the use and the rhetoric of machines in the practice of scoring writing, Ericsson called "troublesome" the position of Thomas Landauer, an AEE software developer and a co-inventor of Latent Semantic Analysis, who in explaining automated scoring remarked that "the fundamental idea is to think of a

paragraph as an equation: its meaning equals a combination of the meanings of its words" (Ericsson, 2006, p. 29). Ericsson pointed out that posted on Landauer's company's website is the claim that its essay scoring engine contains "machine-learning technology that understands the meaning of text" (p. 28). Ericsson is concerned because creators of AEE often claim that machines understand meaning despite the obvious fact that they have, as she notes, "no understanding, no sense of the concepts and the ideas that underlie the words" (p. 32). Or, in the words of Bob Broad (2006), reading and writing involve "complex interpretations of cultural codes" (p. 228). Yet, "evaluation software doesn't even begin or claim to assess these cultural and intellectual capabilities" (p. 228).

Landauer's dual claims move in opposing directions, the first implying a reduction of semantics to syntax (paragraphs are equations) and the second suggesting that these syntactic parsers can somehow grasp semantics (machines understanding meaning). A closer examination of the history of computer science suggests that this rhetorical slippage, the movement to skim off semantic features when talking about computer *processing* but add them back in when talking about computer *output*, has served different communicative functions. At the nascency of computer science, it was necessary to stretch the typical meaning of words in order to promote understanding and describe new problems: Alan Turing wrote of a machine that could "effectively remember," "scan" and "writ[e] down" various symbols (1936, p. 231). Even the word "computer," which now denotes mechanical processing, was borrowed by Turing from the now archaic usage to refer to humans performing mathematical calculations. On the other hand, his contemporary Claude Shannon (1948) framed messages as logical constructs represented in bytes of information in order to solve engineering problems

involved in preserving the physical properties of a message during transmission by telegraph (in Shannon's day) or telephone.

While discursive moves like Turing's and Shannon's followed naturally from the need to stretch language to accommodate technological innovation, the rhetoric of the AEE movement can be seen as a strategic effort to align automated scoring with the computationalism movement, a scientific orientation that sought various forms of unification between minds and machines. In this chapter, I offer a partial historiography of the "computer metaphor" as it has been used to effect what Bazerman (1999) has called the "symbolic engineering" of the earliest automated essay scoring program, Project Essay Grade (PEG). The computer metaphor, the Church-Turing Thesis, and the "Turing test," I suggest here, function both explicitly and implicitly as a rhetorical and conceptual ground of this early work in the field of automated essay evaluation.

Furthermore, early work around AEE has emerged as both a reading and *mis*reading of Alan Turing's contributions to cognitive science via his study of computation. These readings and misreadings of the Church-Turing Thesis and the "Turing test" served as a context in which the computational blended with the mental and, one might speculate, explain some of the antipathy toward AEE among social constructivists. (It is not that AEE should be viewed apart from the material conditions that sustain it and the very real material implications of the possibility of replacing humans with machine labor in the classroom, but that the choice of Turing as a muse gave compositionists an intellectual point of resistance apart from what might be imputed as their personal interests.)

The Computer Metaphor

One of the prevailing dogmas of the brain sciences—a commitment so ubiquitous that it often functions implicitly—is that the human mind is a type of computer. "The guiding idea of cognitive science," reads one introductory textbook, "is that mental operations involve processing information, and hence that we can study how the mind works by studying how information is processed" (Bermúdez, 2010, p. 6). Reads another, "[M]any but not all cognitive scientists view thinking as a kind of computation and use computational metaphors to describe and explain how people solve problems and learn" (Thagard, 2002, pp. 3-4).

The "computer metaphor" that cognitive science embraces is a trope that asserts, as computer scientist Matthias Scheutz (2002) puts it, that "the mind is to the brain as the program is to the hardware" (p. 7). The computer metaphor affixes computer science to the brain sciences, as Scheutz (2002) goes on to say:

> It is this computer metaphor that underwrites the rebirth of computationalism in the twentieth century, and the birth of what is nowadays known as *cognitive science*…: by viewing cognitive functions as computations, explanations of mental processes in terms of programs become scientifically justifiable without having to take neurological underpinnings into account—the 'wetware brain' is simply viewed as a computer on which the software 'mind' is running (or if not mind itself, then at least all the cognitive functions that constitute it). (pp. 8-9)

While this metaphor of mind as computer offers a number of interpretive schema, the claim Scheutz points to maintains that it is possible to talk about minds at multiple conceptual levels, with one of those levels involving some sort of transformation of input to output. (Just as it is possible to talk about computers as both electronic circuit devices and as weather forecasters or chess-players, a mind may be a collection of neurons, a contemplator of astrology, and a symbol manipulator.)

The metaphor of brain as computer might further be considered a *threshold concept*—a new and surprising way of thinking about a concept that despite its novelty is nevertheless imperative to further progression in the discipline (Meyer & Land, 2003). Indeed, the computer metaphor seems to be contained within a set of disciplines studying the brain, placing it at odds with at least some versions of the kind of social constructivist (e.g., Vygotsky, 1978) doctrine that is held by many scholars in composition studies.

Returning to the Ericsson/Landauer narrative and counternarrative, I propose that what troubles Ericsson, at least in part, involves the commitments of computationalism—that mental processing is algorithmic, that computers running the right algorithms can "think"—and are part of a broader cognitive science discourse not limited to AEE but employed by it. Such discourse can be expected to strike social constructivists as incongruous, those who hold that meaning is irreducibly found in the social relations that agents have to one another. At least, this is one version of the tension between constructivist theories of knowledge and the computer metaphor. I will continue to refine these notions in what follows. While it would be beyond our scope to explore the full extent of the relationship between AEE and computational thinking, the rest of this

chapter aims to provide some sketches of how this has played out in the early stages of AEE; namely in Ellis Page's use of Alan Turing's work.

Turing Computability: an Interpretation for Composition Studies

Computationalism, underwritten by the computer metaphor, traces roots to theories of *computability*. The concept of computability emerged during the early 20th century (although its modern-day roots go as far back as Gottfried Leibniz). For our purposes, the critical point of origin lies in the pioneering work of mathematician Alan Turing and its radiation into the fields of computer science, cognitive science, linguistics, and philosophy.

Turing's seminal paper, "On Computable Numbers, with an Application to the Entscheidungsproblem" (1936) concerned itself with a particular problem in mathematics—the decision-problem—of whether there existed an "effective method" of determining whether any given formula belonging to a system of symbolic logic were provable in that system. An effective procedure, or systematic method, amounts roughly and informally to any method which can be carried out by a human computer through finite instructions, in a finite number of steps, requiring no insight, and which will definitely work if no errors are made (see discussion in Copeland, 2004, p. 42). This decision-problem interested Turing's contemporary mathematician David Hilbert, among others, because the creation of a system of mathematical logic in which every mathematical proposition was decidable would have the surprising consequence that one could determine the truth of any mathematical proposition, once one had developed a system of expressing mathematical propositions in formal terms (Copeland, 2004, pp. 46-

47). In other words, if the decision-problem could be solved, mathematical proofs could be more or less cranked out by a mechanical process.

In a proof that is beyond the scope of this chapter, Turing showed that there could be *no* general process for proving any given formula in a system of mathematical logic. The *Entscheidungsproblem* had no solution. Intuition, combined with good old-fashioned toil and sweat, would have to do for discovering mathematical truths. Turing proved that not even the predicate calculus (which contains the logical building blocks of any formal system of arithmetic) is a decidable system—we cannot simply set a computer in motion to tell us whether Fermat's last theorem is true (i.e., provable) and wait for the output. (As Copeland explained, if the predicate calculus is not decidable, then neither is arithmetic (2004, p. 52). This is what Turing's proof showed. In contrast, the propositional calculus *is* decidable because one can construct a truth table revealing whether any statement is true in the system or not).

Turing's result may strike the reader as a technical discovery affecting only the field of number theory, but the process he used for arriving at his result has been generative in many fields and is a root of the "computer metaphor" spoken of at the outset of this chapter. Because of the specific way Turing's proof is framed, he had to create an imaginary mechanical device that was capable of computing any real numbers that could be calculated by means of a rote procedure (Turing 1936, p. 230). Turing then showed that *any* mathematical calculation that could be performed by someone following a rote set of instructions could also be carried out by this device. [3] Turing's proof that the "decision problem" had no solution employed the concept of a computing machine that

[3] See rendering proposed by Copeland (2004): "Any systematic method can be carried out by the universal Turing machine" (p.41).

could compute anything that was the result of an effective procedure. Because Turing's contemporary Alonzo Church had arrived at another method for showing the same result, Turing's theoretical result became known as the Church-Turing Thesis.

To get a better picture of this result, imagine a machine capable only of printing ones and zeros. A Turing machine instructed to count from ten to 20 (in binary code, say) could carry out these instructions by following a set of commands such as "if current state reads *zero*, erase it, print *one* and move one cell to the left." (In today's parlance, the set of commands becomes the *program* for the machine.) Note that a machine counting from ten to 20 by ones is actually an adding machine, so Turing's machine can perform arithmetic by simply changing numbers from one to another. The kind of machine that Turing conceived of, a universal computing machine—now typically called a Universal Turing Machine—is achieved in abstract terms by encoding a program in numeric form (a "description number") onto the tape that the computer uses to perform its computations and produce its output. Digital computers of today are examples of Universal Turing Machines, because they can carry out any mechanical set of instructions.

In other words, and this is the important point for our purposes—the Church-Turing Thesis connected systematicity with mechanization (see Copeland, 2004). If some process can be expressed via a formula with an effective procedure to carry it out, that process could be computed by a mechanized instrument. Even procedures that humans carry out "intuitively," such as multiplying or dividing, can be broken down into steps that involve no mathematical knowledge whatsoever, but simply the writing and erasing of various symbols. The Turing machine transforms math into its own atomic steps of moving, printing, changing state, and erasing.

To say that a computer is "calculating," then, is as much a metaphor as to say that a computer is "reading" or "writing." Human "computers" were well-known to Turing, as they were fundamental to the war effort by performing numerical integration and other mathematical operations in order to compute missile trajectories and such. So Turing was concerning himself with human calculation, that symbolic, meaning-driven activity, and showing that if this activity was rotely specifiable in a certain way, a mechanical device could carry it out.

Expanding this point, it seems to me that both opponents and proponents of AEE reductively mischaracterize it as a transformation of words into numbers. But computer technology does not represent a pitting of mathematical practices against literacy practices: it represents the transformation of *computable meaning-making activities into mechanical action*. It just so happens that mathematical calculations are the sort of meaning-making activities that easily lend themselves to the kind of algorithmic specificity required by computing devices. As Copeland (2004) writes, "if the Church-Turing thesis is true, then talk about the existence or non-existence of systematic methods can be replaced throughout mathematics and logic by talk about the existence or non-existence of Turing machine programmes" (p. 43). Seen through the lens of the Turing machine, AEE involves reducing one set of symbols found in natural language to a set of symbols found in computer science.

In this way, the cyborg—the coders and the machines performing these transformations—is *writing* symbols, both literally and metaphorically. Haraway's reflections are apt here, "[T]he silicon chip is a surface for writing; it is etched in molecular scales disturbed only by atomic noise, the ultimate interference for nuclear

scores" (p. 280). Inscribed on its surface are the electronic pathways that enable this other form of writing, the illumination of ones and zeros which, translated back up into various levels of code, equals an automated essay score or the sending and receiving of an email message.

From Computability to Computationalism

While the Church-Turing Thesis said that if something is a mathematical function calculable by "finite means" (Turing, 1936, p. 230) it can be instantiated mechanically, the commercialization of computers and the advance of computer programming invited innovators to broaden the class of phenomena that could be represented mathematically and thereby calculable by computer. Among other phenomena, brain activity came to be studied computationally. Indeed, even as Turing's result showed that the decision-problem was not provable and thus the field of mathematics could not be reduced to mechanical action alone, his Church-Turing Thesis paved the way for the proliferation of the "computer metaphor" of mind. This, as we will see, gave AEE founder Ellis Page a way to legitimate his new enterprise. And Turing himself seemed at times to advocate for the search for computational approaches to understanding. So let us now turn from computability to computationalism.

In 1994, philosopher Patricia Churchland and biologist Terrence Sejnowski published the book *The Computational Brain*, in which they laid out an extended case for a computational approach to the brain called "computational neuroscience." Churchland and Sejnowski argued not only that computational approaches can serve as a research tool in modeling "complex systems such as networks, ganglia, and brains" (p. 7), but also

that, more strongly, the brain itself is a form of computer, that "what is being modeled by a computer is itself a form of computer" (ibid).

Churchland and Sejnowski's premise was not unusual for mid- to late twentieth century thinking in cognitive science. Whether the mind was a computer, whether computers were themselves minds, and whether the "computer metaphor" was explanatory of cognition were debated contentiously during this time. In a famous counterargument that came to be known as the Chinese room thought experiment, philosopher John Searle (1980) argued that a computer responding to English questions was no different from an English speaker responding to questions written in Chinese by following a set of instructions written in English, indicating which characters to associate with which questions. (E.g., the Chinese room would have instructions such as, "if the questioner says, 'ni haoma?' respond with 'wo henhao'.) Such a computer would not understand Chinese, nor would its actions explain how humans understand languages (pp. 285-286). Again, the contested problem of meaning looms large in Searle's objection.

We can see that Churchland and Sejnowski's position falls under the computer metaphor advanced earlier. On this doctrine, the brain is itself a computer, reflecting the odd realization that computers are themselves modeling a kind of computer, "albeit one quite unlike the serial, digital machines on which computer science cut its teeth" (Churchland & Sejnowski, 1994, p. 7).

Turing's biographer Jack Copeland has claimed that the assertion of the computer metaphor of mind as established doctrine, rather than as scientific hypothesis, rests on a confusion over the scope of the Church-Turing Thesis. Wrote Copeland:

> [A]n error which, unfortunately, is common in modern writing on computability and the brain is to hold that Turing's results somehow entail that the brain, and indeed any biological or physical system whatever, can be simulated by a Turing machine….The Church-Turing thesis does not entail that the brain (or the mind, or consciousness) can be modelled by a Turing machine program, not even in conjunction with the belief that the brain (or mind, etc.) is scientifically explicable, or exhibits a systematic pattern of responses to the environment, or is 'rule-governed' (etc.)….[I]t is an open question whether a completed neuroscience will employ functions that are not effectively calculable." (2002, "Misunderstandings of the Thesis")

Copeland is saying that the Church-Turing thesis does not imply that *any* rule-governed function whatsoever can be computed by a Universal Turing machine or its equivalent, a digital computer, because there is indeed a class of functions that are not computable. Piccinini (2007) has also argued that only a completed neuroscience will tell whether, and in which ways, mental states are Turing-computational, and therefore whether computational realism turns out to be true. Writes Piccinini: "CTT [i.e., the Church-Turing thesis] pertains to functions that are computable in the intuitive sense employed in mathematics. In order to show that something falls under CTT, it must first be shown that it is computable in that sense. CTT, per se, does nothing to establish that something is computable. Because of this, supposing that CTT entails computationalism is a fallacy" (2007, p. 99). Ordered pairs that are put together at random, for instance,

would be part of a noncomputable function as would functions that use continuous as opposed to discrete variables, Churchland & Sejnowski, 1994, p. 62; Piccinini, 2007, pp. 100-104). Not all deterministic systems (ones in which the future states depend on present states) are algorithmically computable systems (Penrose, 1999, p. 220).

The implications of uncritically assuming that brains instantiate computational states (committing the CTT fallacy) would be an appealing move for an AEE proponent, and indeed I explore this in the following sections. The implication would be this: if cognitive states turn out to be computable states, then the cognitive act of grading an essay is ultimately itself computable. Successfully automating essay scoring then becomes a question of specifying the right functions that articulate the cognitive act of grading so that the computer can replicate this act. Thus far, developers have presumed that the functions being computed by human graders have solely to do with analyzing the properties of the essay itself, an assumption that is of course deeply contested by compositionists who subscribe to the social theory of composing. Ellis Page, as I will try to show, incorporates assumptions about the computer metaphor as a rhetorical strategy.

But the assumption that brains simple are computers is not the only stance that could be a potential driver for AEE. Piccinini (2007) goes on to distinguish processes or calculations that are truly Turing-computable from those that can be mimicked by computer algorithms. He points out that much of the attraction of computational approaches in the sciences is not that biological or physical systems *are* computational systems but that they can be *simulated by* computational systems even when these processes cannot be instantiated computationally. In the following passage, Piccinini observes the utility of simulation rather than replication:

> The modern study of dynamical systems has exploded over the last half-century, to a large extent, thanks to the advent of digital computers. This is because computers, by offering larger and larger amounts of memory and computation speed, allow scientists to develop methods of approximation for systems of equations that are not analytically solvable, so as to study their behavior based on those approximations. (2007, p. 105)

As Piccinini goes on to observe, claiming that certain mental activities can be simulated or approximated by computer algorithms does *not* invoke a misreading of the Church-Turing thesis. Certainly computers can, and do, approximate many natural processes, such as the weather, without these processes being strictly Turing computable. Approximation is, then, an entirely different construct from computability.

Formulated as an attempt to simulate human grading rather than Turing-replicate it, AEE is a more realistic endeavor. Fundamentally—and I take this to be accepted by AEE proponents and detractors alike although it is not much discussed in the literature—in order to Turing-replicate human scoring, one would need *a computational theory of reading*, and more specifically, a *computational theory of evaluation*. Such theories of reading are developing, but they are not complete. These theories would not, however, only involve the kind of textual feature analysis that AEE can offer; instead, according to Mason, Tornatora, and Pluchino (2013), most models of reading assume that, at the abstract level, the representation involves the union of both the prior knowledge of the reader and new information from the text. "The common issue shared by all models is the

constructive role of the reader who activates her or his relevant background knowledge as she or he processes text information, and tries to integrate background knowledge with the new information to build a coherent representation of the text, that is, the situational model in Kintsch's (1998) terms" (Sec. 1.1). So, if computational reading experts are to be believed, AEE cannot replicate human scoring until, at the very least, it can model or simulate the extratextual component of reading.

I am sure that AEE proponents would grant that current AEE has not identified effective procedures that would enable a computer to replicate evaluating an essay. Yet, the history of AEE does in many senses still begin with the history of the computer, of Turing machines. The developers of AEE, particularly Ellis Page, called upon Turing and the vision of Turing-replication as a rhetorical maneuver to legitimate their emerging and surely shocking technology to prospective audiences. In so doing, I maintain, they committed the Church-Turing fallacy of assuming that any process that is systematic can be replicated mechanically.

Charles Bazerman's (1999) concept of "symbolic engineering" helps frame what the founders of AEE may have been doing as they sought to promote and legitimate a new technology. For Bazerman, symbolic engineering is the development of "symbols that will give presence, meaning, and value to a technological object or process within a discursive system" (p. 335). Bazerman contended that most technologies do not simply speak for themselves, are not simply themselves sufficient to constitute their acceptance into society. Most technologies need symbolic mediators that mobilize the heterogeneous values of different aspects of society. Bazerman's concept of symbolic engineering is a lens to understand how Ellis Page and his co-founders sought to represent automated

scoring to various audiences did not simply function descriptively (i.e., to represent the state of the technology) but aspirationally, as attempts to further its material success.

To outline the argument to be presented, I explore the ways in which Turing's status in the early, promise-filled days of computing lends Page's project the value that Bazerman describes as credibility (p. 338), which is important in introducing a technology into a public forum or, in Page's case, to other educators. While in Bazerman's extended example, Thomas Edison used himself to lend credibility to his endeavors, Turing was a figure so luminescent in the field of computing that it is entirely reasonable that Page would have turned to him to establish an ethos. Turing's "Test" of intelligence is Page's principal tool of symbolic engineering. A brief description of Turing's proposal follows.

Turing's "Test": The Imitation Game

Alan Turing had originally concerned himself with mechanizing the computation of mathematical propositions, but perhaps his most popularly familiar contribution to the study of intelligence came in a later non-mathematical writing. As he put it in this seminal work, "Computing Machinery and Intelligence": Can machines think? (Turing, 1950).

Turing's essay describes a thought experiment he called the "imitation game," wherein an interrogator would try to determine whether a mystery respondent (hidden behind a screen and printing out answers on a printer) were a person, or a computer. Turing stipulated that the object of the first respondent, the computer, was to cause the interrogator to guess wrong, thus responding as humanly as possible. The second role-

player, the human, had the object of leading the interrogator in the right direction, in other words, also answering as humanly as possible (p. 434). Would the interrogator decide correctly between the two as frequently as he did if, say, he were trying to determine which of the two role-players were male and which were female?

Because Turing was aware of the nascency of computer science, he took pains to explain to his readership the kind of mechanical device he had in mind that could play the imitation game, and to provide an explanation for its potential success. The machines in Turing's example were digital computers, and as such, were able to carry out any operations that can be performed by a "human computer."[4] A human computer, he says, is someone charged with following "fixed rules" from which he or she is not allowed to deviate (a reformulation of Turing's earlier statements on computability.)

Turing predicted that in 50 years from the time of his 1950 writing, computers would be able to mimic human responders so accurately that in the imitation game, "an average interrogator will not have more than 70 percent chance of making the right identification after five minutes of questioning" (442). As Saygin, Cikekli and Akman (2000) note, Turing's prediction failed: no machine had been able to pass the Turing test by the turn of the millennium. Another way of framing Turing's prediction would be that

[4] Turing describes a digital computer as having three parts: store, executive unit, and control. The store compares to what the human computer would use to execute instructions, including human memory, the pages of the instruction book, and a paper for calculations. The executive unit carries out the individual operations that make up any given calculation, and the control is constructed so that the calculations are carried out properly (p. 437). For instance, a number such as 6809430217 could be an instructional code for "add the number stored in 6809 to the number stored in 4302" where the number 17 stands for the instruction to "add" (p. 437). For those unfamiliar with the nomenclature of early digital computing, imagine that software programs constitute instructions, that the CPU carries out those instructions and that the memory is where the instructions and the calculations are stored. So again, the computer is working by rote.

the computer would be so good at mimicking the human that the interrogator would guess wrongly in at least three out of ten cases, making the Turing test a success. Conversely, the computer will be able to foil the human interrogator at least 30 percent of the time. As we will see, this idea of probabilistic mimicry factors into the original (and many current) justifications of AEE.

Table 1 presents my rendition of a side-by-side comparison of three possible outcomes of ten imitation game trials: a computer that never fools the interrogator (A); a computer that fools the interrogator 30 percent of the time (B); and a computer that fools the interrogator half the time—an indistinguishable result (C). Turing's prediction is conservative. He allows room for the computer to be distinguishable at least three out of ten times. His "game" is thus vulnerable to the rebuttal that it is not strong enough: in order for a machine to be considered intelligent, a guesser should have a record no better than chance.

Conversation number	Scenario A: Interrogator always guesses correctly	Scenario B: Turing's prediction	Scenario C: Computer fools interrogator
1	Correct	**Correct**	Correct
2	Correct	**Correct**	Incorrect
3	Correct	**Incorrect**	Correct
4	Correct	**Correct**	Correct
5	Correct	**Incorrect**	Incorrect

6	Correct	**Correct**	Correct
7	Correct	**Correct**	Incorrect
8	Correct	**Correct**	Correct
9	Correct	**Correct**	Incorrect
10	Correct	**Incorrect**	Incorrect

Table 1. Imitation game outcomes.

Let us take stock of two common objections to Turing's postulation. The first objection is that language is not Turing-computable; therefore, there is no way that a machine could replicate human language even 30 percent of the time, let alone more frequently than that. Turing does consider this objection (as he considers many others) that humans are not bound by rules for every set of circumstances—natural language is not "computed" in the manner of mathematics—but his response seems to be that we cannot *prove* this is so. He writes, "the only way we know of for finding such laws is scientific observation, and we certainly know of no circumstances under which we could say, 'We have searched enough. There are no such laws'" (p. 452). The implied warrant here is that without such laws, machines would not pass the Turing test; the test is not a test of approximation but of replication.

Seemingly, for Turing it is an open question whether rules will be discovered for all or most aspects of human behavior; for as Piccinini (2007) writes: "[T]he consensus among Turing scholars is that in arguing for [the Church-Turing thesis], all that Turing was attempting to establish is that human computation processes are computable by [Turing machines]—he was not trying to establish that all mental processes are the

following of an effective procedure" (p. 109). In other words, if a process is an act of human computation it can be machine-replicated; the question becomes which human acts are acts of computation. And what Turing points out is that this objection pre-empts the response that natural language may have such rules even if such rules have not yet been uncovered.

A second objection Turing calls the argument from consciousness. This is the argument that machines are not equivalent to brains because machines do not have thoughts and feelings. Turing quotes the early artificial intelligence skeptic and surgeon Geoffrey Jefferson to flesh out this objection: "Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain" (Jefferson, qtd. in Turing, 1950, p. 445).

Turing answers this objection by reciting a version of what philosophers call the problem of other minds: we actually have no direct, experiential proof that other *people* think either, so for that matter, how can we know if anyone is conscious? Rather than accept such a solipsistic point of view, Turing points out that we ascribe thoughts and feelings to others, at least if we seem to be able to converse with them and they seem to be answering in *viva voce*, and asks essentially why the standard should not hold for machines as well. The problem of consciousness intersects most with the discourse surrounding AEE, and we will return to it anew in the following section.

As we consider the viability, and the liability, of grounding a new technology in Alan Turing's credibility and the legitimacy of his measure of machine intelligence, a disagreement in how to interpret Turing's intentions needs to be raised. It seems that

Turing further held that the full Turing-computability of language was necessary for the test to be passed by a computer (given his reference to discovering laws of human behavior). Surely, this would seem to be an ambitious hope. It is plausible to think that a computer that was running an approximation of linguistic abilities might pass the test. Indeed, as Piccinini has noted, replication represents a higher standard than simulation or mimicry; a computer that could be detected by an interrogator with even 30 percent accuracy is probably not Turing-replicating human responses but simulating them, with limited accuracy. In turning to consider Ellis Page and his Project Essay Grade (PEG) software, one case see that his readings and misreadings of Turing represent a rhetorical attempt to symbolically engineer his technology.

Is PEG an Approximator or a Replicator?

In making the case for Project Essay Grade in his influential "The Imminence of…Grading Essays by Computer" (1966), Ellis Page employed the rhetorical strategy of arguing, both implicitly and explicitly, that PEG passes a Turing test. Implicitly, he did so by presenting his readers with a correlation matrix headlined by the question, "Which One is the Computer?" The matrix represented what he said was an agreement intercorrelation between five judges of 138 student essays, correlations which ranged from .44 to .61. One of the judges was Page's software, rather than a person. "To display the 'humanness' of this achievement," Page wrote of the matrix, "we performed an extremely conservative cross-validation" (p. 241). Each number in the matrix represents a measure of correlation between two of the judges' scores (judges are listed as A-E).

Figure 1 reproduces the correlation matrix below, preserving Page's title for rhetorical effect. The implication of the matrix is that the computer "judge" is indistinguishable from the human judges on the basis of the agreement statistics. Clearly some questions could be raised about the vagueness of Page's description of a cross-validation without specifying the statistical test used. For now, though, other rhetorical choices are of interest.[5]

TABLE I.
WHICH ONE IS THE COMPUTER?

Below is the intercorrelation matrix generated by the cross-validation of PEG I in the following fashion:

All "judges" graded the overall quality of a set of 138 essays written by high school students in grades 8-12. One "judge" was a computer, the other four were independent human experts. The correlations in this table show the extent to which each "judge" tended to agree with each other in grading essays. The computer-assigned grades were based upon beta weightings generated from the multiple prediction of human judgments on 138 essays by *other* students randomly drawn from the same population. Which one, A, B, C, D, or E, is the computer vector?

	Judges				
	A	*B*	*C*	*D*	*E*
A		51	51	44	57
B	51		53	56	61
C	51	53		48	49
D	44	56	48		59
E	57	61	49	59	

Figure 1. Intercorrelation matrix provided in Page (1966).

[5] Indeed, it is not clear how Page's contemporaries received his work. The only response to Page from his contemporaries that I have found is an article by Levy and Fritz (1972) entitled, "Status Report on Computer Grading of Essays." There may have been other responses but what has primarily survived in the literature is Page's work itself.

At the end of his discussion Page does not, after all, reveal which judge was the computer. The object of the exercise, then, could not have been to communicate any findings of the *actual* correlation between the human and the computer agreement. Instead, I would argue, his objective was to set up his readership in the position of interrogator, then showing how such readers could not distinguish a computer disguised as a human judge. In other words, Page set up a Turing test of his own, connecting the replication of interrater agreement correlations with passing the test. In so doing he connects AEE with a symbol of the then-exploding field of Artificial Intelligence.

Furthermore, continuing a trope from Turing, Page predicted that the one way to eventually distinguish computer from human raters is that the former will actually correlate *more* highly with the human raters than the humans themselves will, due to the elimination of random error found in human rating (p. 242). Here, Page adopted another point that Turing made: the human reveals him- or herself by the propensity for making mistakes. "If the man were to try and pretend to be the machine," Turing wrote, "he would clearly make a very poor showing. He would be given away at once by slowness and inaccuracy in arithmetic" (1950, p. 435).

But Page's invocation of Turing did not stop with oblique implications that the software has passed the Turing test—a formulation that, as I have argued above, potentially conflates simulation or approximation of human activity with a strict Turing computability. Page (1968) appealed to Turing directly to justify his presentation of the same data set presented in his first essay: "we feel that A.M. Turing, who recommended the 'difference test' as a good trial of the presumably intelligent machine, might well be pleased" (p. 217). And then, Page goes on to suggest the much stronger and extremely

surprising claim that automated essay grading is indeed *an instantiation of a Turing computable process*:

> From other articles in this issue it must be apparent that the computer is a natural agent in the analysis of language. When considered physically, after all, a student essay is, regardless of its subject matter or the standards of its evaluation, merely a set of recurrent symbols arranged in a linear string. And the computer is, of course, the symbol manipulator without peer, of superhuman speed and accuracy. Thus the problem of evaluating an essay may be regarded as the problem of transforming a string of input symbols into some appropriate string of output symbols. And such a string problem calls for a "Turing machine" which, by reading and acting upon just one symbol at a time, can achieve any arbitrary level of complex response which might be practically necessary. In such a model, the input string could consist of the sequence of letters in the student essay, and the output string could consist of a set of printed grades, or comments, or diagnostic profiles. (Page, 1968, p. 211)

While it is accurate to say that a Turing machine would transform a string of symbols into another string of symbols, it cannot achieve any "arbitrary level of complex response" if that is to mean that the output string resulted from computing any function whatsoever. (That is why the Church-Turing thesis talks of computable functions, not simply any functions.)

But more pressingly, Page seems so eager to cast automated scoring as a computer innovation ("after all, a student essay is…merely a set of recurrent symbols") in a way that prefigures Landauer's paragraph-as-an-equation approach, that what is totally elided is the problem of choosing the right algorithm to be applied to that string. Here, Page's conclusion seems to commit the Church-Turing fallacy. In saying that the "problem of evaluating an essay may be regarded as the problem of transforming a string of input symbols into [a] string output symbols," he seems to presuppose that scoring is computational, that human graders carry out effective procedures (algorithms). The fact that multiple, trained, human graders can come up with different scores is evidence that such an algorithm, if it exists even in theory, is much more complicated than anything previously concocted. It is even hard to imagine what a computational approach to grading would mean.

Page's use of Turing, instead, could be read as an attempt to give PEG a meaning that arched toward the hope that computers could perform communicative tasks much like people do, the promise of passing a Turing test (C. Bazerman, personal communication, August 1, 2014). Page's way of creating his Turing-style test served to mask the computer's performance assigning individual scores and emphasize score averages, a theme that will be picked up and questioned in subsequent chapters.

Jeremiah Dyehouse (2007) has explored this slippage from within another framework, and his work also points to the way in which Page's rhetorical choices were efforts to symbolically engineer the technology. In an essay emphasizing the role of argument in stabilization of technology products (as a complement to Bazerman's emphasis on discursive influence), Dyehouse showed how Page initially framed his work

as a "ratings simulation" in antithesis to the kind of "master analysis" of writing that a human could carry out based on intrinsic qualities. Still, Page suggested that someday computers may be able to reach the point of master analysis (Dyehouse, 2007, p. 123). Dyehouse's rhetorical analyses demonstrated that Page was progressively able to characterize his work as a simulation with the idea of a master analysis eventually fading into the background as computational applications in various fields of knowledge production gained traction (p. 132). What Dyehouse's analysis implies is that Turing computation—which implies the existence of a master analysis—was more of a rhetorical framing device than an actual commitment of Page's.

In their (1995) study, Page and Petersen summoned Turing again, in a response to what they call the "humanist" objection to AEE, which they deem to mean rejecting AEE out of hand because its lack of sentience prevents it from measuring in the way that people do. The humanist objection can be compared to Turing's argument from consciousness, which placed sentience as a central marker of intelligence. Here are Page and Petersen:

> Some of the humanist objections hark back to the dawn of the computer revolution. It was asserted that certain choices require human knowledge and background wisdom, while the computer will do only what it is programmed to do. It will not understand or appreciate an essay, and so it cannot measure what a human judge measures. Its judgments should be rejected out of hand.

> Such arguments were so common 40 years ago that Alan Turing devised a response in the form of his classic "difference game," now widely known as the "Turing test." Imagine that you have a person behind one door and a computer behind another and that you don't know which is which. You are allowed to slip notes under the door, read the printed responses, and try to determine which door hides the computer. If you are unable to find a relevant question that will reveal which is the machine, the computer wins the game.
>
> Suppose we play another version of this game. You have seven doors with human raters behind six of them and a computer behind the seventh. We pass essays under all seven doors and get back a score (or set of scores, if we're rating traits). We examine the scores and continue to pass essays and collect ratings. Can we tell which door hides the computer? When we study the results with 300 essays, we find -- lo and behold -- that we can easily identify the computer. It's the one that agrees best with the other judges! Where does this leave the humanist objections? (Page & Petersen, 1995, Section 9, paras. 2, 3 & 4)

For Page and Petersen, passing a Turing test amounts to overcoming the humanist objection, presumably because sentience would no longer be seen as necessary to perform a task like humans do. Notice, however, that a certain version of the humanist objection simply rules out difference tests *tout court*, because the salient matter is having

a mental life, not behaving as such. Turing notices this while Page does not; Turing says of the argument from consciousness that it "appears to be a denial of the validity of our test" (1950, p. 446), which is why he turns to the problem of other minds in response. The main point for our purposes is to notice the move to frame computer scoring as a version of a Turing test.

In his (2003) essay, Page added the following emphatic closure to the paragraph immediately above: "PEG [Page's software] has shown the world one solution to the Turing Test" (p. 52). Page's modified reading of Turing's test suggests that the computer is actually a *better* mimic of humans than humans are of each other. What Page and Petersen meant statistically in the case in question is that the computer "judge" did a better job of predicting the average score of six human judges reviewing a set of 300 essays, than two of the human judges did in predicting the same average score (1995, Section 2).

By the 1960s, Turing's "test" had generated an appreciable amount of attention, and a number of philosophically-oriented replies had emerged, some fortifying his argument and others challenging it (for further discussion of replies to Turing, see Saygin, Cikekli & Akman, 2000). However, Page's rhetorical appeal in his (1968) essay seemingly represents neither critique nor extension of Turing's test but rhetorical use of *ethos*: an appeal to authority (e.g., proposing that Turing "might well be pleased") at the same time as it represents an attempt to legitimate the litmus test that Page has set up for his cyborg reader. As Bazerman has noted, a technology must not only be able to "succeed materially…it must also succeed symbolically" (p. 335), which might be said to

mean that the technology has to make itself culturally acceptable not only functionally adequate.

But, Page's efforts to legitimize PEG point to the interconnection between the functional and the symbolic—what it meant to succeed materially as an essay grader was as much architected by Page through his use of the correlation matrix version of the Turing test as what it meant to gain acceptance of relevant discursive communities. He was arguing for the legitimacy of the test as well as the technology. The test of whether Page's technology works—perhaps unlike in Bazerman's extended example of Thomas Edison whose product's functionality was universally manifest (by turning the thing on, so to speak) depends on the degree of acceptance of the particular measure of success to which the technology is subjected. In other words, in cases such as automated scoring, the meaning of the technology "working" depends on establishing the legitimacy of a certain test—Page, of course, turns to Turing to validate his version of the "difference test."

As additional evidence that Turing's difference test methodology has been used to legitimate automated scoring, here is ETS researcher Yigal Attali (2013) writing in the *Handbook of Automated Essay Evaluation*: "most validation efforts of [AEE] have implicitly or explicitly tried to answer [the humanist] objection by showing that machine scores are indistinguishable from human scores….Page and Petersen (1995) even suggest a version of the Turing test…*to emphasize the importance of score imitation for the acceptability of [AEE]"* (Attali, 2013, p. 182). Unlike Page, Attali realizes that score similarity in itself does not adequately respond to the humanist objection if this objection is truly about sentience, but we will come to this matter later on.

The Turing test has remained an aspect of the symbolic engineering of AEE, right up to the 2012 Hewlett Foundation-sponsored competition for automated scoring vendors, which Hewlett program director Barbara Chow billed as "a single place that works as a Turing test. Our hope is to provide a neutral, impartial platform on which to make judgments of whether automated scoring can match human raters" (Edsurge, 2012). The promise here is not that computers be *sentient* intelligent beings, or care about the essays they grade, but that they have similar interrater agreement patterns as human raters.

The symbolic appeal to Turing was, obviously, not without its baggage. The measure that Page chose to give meaning to his new technology, the Turing test, seems to alienate the composition community, those who would ultimately need to adopt it for it to gain material success. (Indistinguishability tests were not new with Turing; rather, they date at least to Descartes, and as many have pointed out, Turing's test invokes a simplistic equivalence of output as its criterion.) Turing's test effectively equated intelligence with the ability to mimic the linguistic structures and functioning of natural language to a certain degree of accuracy (guessing correctly only 70 percent of the time after five minutes of the imitation game). Failing to make sentience a condition of intelligence invites, as shown previously, the humanist objection.

Also, Page's formulation contains at least one notable and seemingly problematic departure from Turing's. The departure is that in the case of automated scoring, the interrogator is only privy to the responses of the subjects, but is not privy to the questions. In Page's setup, the interrogator only gets to see the scores themselves, not the actual essays that were scored. Page's version makes only the percent agreement or

correlation visible. Without analyzing individual essay-score patterns, an interrogator would be much more likely to be fooled than if she could examine actual essays along with the scores.

Exploring splits—those essays where the computer score diverges from the human score by more than one adjacent interval—is a possible way to differentiate human from machine scoring. Without looking at the level of the individual essay, we cannot conclude that the Turing test would have been passed; this would be like asking an interrogator to judge between human and computer respondent, without looking at the questions.

Les Perelman has pursued this objection to good effect, recently by using an automated writer called the babel generator (2014, personal communication), which produces essays that humans would recognize as gibberish, but which nevertheless attain high scores from machine graders due to their ability to put together grammatically complex sentences with infrequently-used words (which the computer uses as a proxy for sophistication of prose). [6] The babel generator can be found at http://babel-generator.herokuapp.com/.

But one does not need the babel generator to expose the differential scoring practices of current automated technologies versus human technologies. In fact, the five paragraphs I have just written, beginning with "Quite possibly…," and including the present paragraph were recently fed into a commercial automated scoring engine. The prompt they responded to asked students to consider whether liberal arts education is worthwhile. The four paragraphs above, even though they have no part in addressing the

[6] Although the "Sokal affair" suggests that humans can be fooled by lofty prose, too (Scott, 1996).

question of the validity of liberal arts education, attained a score of *five out of six* possible points from the machine and developed its ideas "with reasons and examples that are appropriate" according to the automated feedback. Any person scoring the previous four paragraphs would have flagged the essay as off-topic and incomplete.

So, providing sophisticated but meaningless or irrelevant prose will almost instantly point out which scorer is the machine. And pursuing this strategy is no less valid a strategy than the interrogator asking the machine to write a sonnet. "Count me out on this one. I never could write poetry," Turing's respondent replies (p. 434). Because machines are fooled by syntactic but meaningless prose, an interrogator could easily craft essays that would betray it. While computer scoring may fare well when scores are examined in the aggregate, a correlation matrix is not the only form of Turing test that could be designed; one that revealed scores and essays would be more faithful to the original conception and more relevant to the concerns of writing studies.

What Page calls the humanist objection (or the argument from consciousness that Turing considers) rejects the very premise that passing a Turing test is enough to validate a technology. The humanist objection makes sentience a wedge between humans and machines, thereby rejecting the computer metaphor outright and embracing meaning and social interaction as irreducible aspects of reading and writing. In the final section of this chapter, we turn to the humanist objection once more.

The Humanist Objection Revisited

The humanist objection has laid down tracks in multiple fields of inquiry, not only in the field of educational measurement. It was employed, ironically perhaps, in the very

field of mathematical logic—to the very question that Turing himself set about to answer: the Entscheidungsproblem. As Copeland (2004) noted, an opposite result for the decision-problem than the one Turing produced would have had a shuttering effect on the discipline. Wrote John von Neumann in 1927, "if undecidability were to fail then mathematics, in today's sense, would cease to exist; its place would be taken by a completely mechanical rule, with the aid of which any man [sic] would be able to decide, of any given statement, whether the statement can be proven or not" (qtd. in Copeland 2004, p. 53) As G. H. Hardy said, "if there were…a mechanical set of rules for the solution of all mathematical problems…our activities as mathematicians would come to an end" (qtd. in Copeland 2004, p. 53).

Three-quarters of a century ago, therefore, one finds mathematicians taking pains to prove that mathematical theorems were not decidable through any kind of mechanical process. Turing, as noted, found an "in principle" argument for mathematics not being reducible to computer science. Mathematicians discuss the philosophical importance of Turing's result to the present day. Roger Penrose articulated this sentiment in his influential book *The Emperor's New Mind* (1999), when he discusses Hilbert and his "hope": "for any string of symbols representing a mathematical proposition, say *P*, one should be able to prove either *P* or ~*P*,…if Hilbert's hope could be realized, this would even enable us to dispense with worrying about what the propositions mean altogether!" (Penrose, 1999, p. 137). Continuing on, drawing heavily on the incompleteness proof of Gödel and Turing's work on the decision-problem, Penrose wrote:

> The point of view that one can dispense with the meanings of mathematical statements, regarding them as nothing but strings of symbols in some formal mathematical system, is the mathematical standpoint of *formalism*. Some people like this idea, whereby mathematics becomes a kind of 'meaningless game.' It is not an idea that appeals to me, however. It is indeed 'meaning'—not blind algorithmic computation—that gives mathematics its substance. (p. 137)

The privileging of meaning over formalism as integral to mathematical insight can be seen as one version of the humanist objection, one version of the argument that there is something special in how humans go about solving problems. Penrose noted the sense of curious justice in the fact that it was Turing's proof that helped contribute to the downfall of Hilbert's program: "There is perhaps some irony in the fact that this aspect of Turing's own work may now indirectly provide us with a possible loophole to his own viewpoint concerning the nature of mental phenomena" (p. 46). Turing, who appeared to hold out the hope that the mental was computational, had already proven that mathematical proofs could not be produced through computation alone.

While Page typically described the humanist objection without reference to any specific detractor, this objection has been raised forcefully by contemporary compositionists (as noted by Attali, 2013) and we can see that Patricia Ericsson's remarks presented at the beginning of the chapter attack a formalist, textually-dependent interpretation of meaning. Further into her essay, Ericsson called upon a number of theorists from various disciplines who espoused some version of the social construction

of meaning to attack AEE. For instance, wrote Ericsson, "[T]he machine has no understanding, no sense of the concepts and ideas that underlie the words, no ability to bring to the words what Berthoff claims is important in discerning meaning—'what we [humans] presuppose and analyze and conjecture and conclude'—all of this adding up to what a human sense of a text might mean (p. 43)" (Ericsson, 2006, p. 32).

Later, Ericsson went on to say that, "Writing for an asocial machine that 'understands' a text only as an equation of word + word + word strikes a death blow to the understanding of writing and composing as a meaning-making activity" (2003, p. 37). In Ericsson's view, writing cannot be a process of creating shared understandings between reader and writer if the reader is just an automaton; writing has been removed from its social context and degraded into a "dumbed down version of writing and composition" (p. 37).

Several of Ericsson's contemporaries have advanced the humanist objection as well. Herrington and Moran (2012)'s version of the humanist argument called upon a position staked by Ed White that "writing for nobody is not writing at all" (1969, p. 167, qtd. in Herrington & Moran, 2012) to emphasize their contention that mechanical writing assignments and mechanical scoring algorithms have a dehumanizing effect on writers. "We can support the humanity of our students as writers by insisting on our own humanity as readers" (1969, p. 168, qtd. in Herrington & Moran, 2012). Herrington and Moran echo White in suggesting that writing to machines is writing for nobody; therefore, writing to machines is not writing at all. Writing to a machine "distorts the very nature of writing itself," these authors say (Herrington & Moran, p. 15). Discussing ETS's *e-rater* automated assessment engine, Les Perelman (2012a) wrote: "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" (p. 125).

As I have tried to emphasize, the force of the humanist objection lies in pointing out the duplicity in claims that at once insist that machines can understand language and at the same time concede that they are just symbol manipulators. Passage of a Turing test was once used to make sense of this *aporia,* held out as a standard by which machines could be said to think. But as I have tried to show, whatever the merits of Turing's test as a measure of intelligence, Page did not appear to set up a true Turing test and therefore AEE does not pass it. Passing the Turing test seems to be a higher bar than originally believed.

So where does that leave us? It leaves us with a test for AEE's legitimacy that is somewhat one-sided because it is statistical in nature and leaves aside examination of how the computer scores individual essays. Subsequent chapters will amplify and spell out this concern more closely by beginning to analyze what happens at the level of individual essays, when a set of essays is scored by humans and by a machine scorer.

I suspect, unlike Ericsson but with Perelman, that the problem with machines not understanding meaning is more of a practical impediment than an *a priori* prohibition. If a machine *could* be built to simulate semantic processing, such a machine may be able to validly assess essays. While more advanced systems will evolve, most likely "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" (2012a, p. 129).[7]

Until that day arrives, Page's assertion, as part of the symbolic engineering of Project Essay Grade, that it or any other AEE system has passed a Turing test is not entirely accurate. Interrogating this claim calls attention to the fact that the standard of interrater reliability is not the only bar by which an AEE system ought to be measured, even in circumstances where statistical reliability is quite high. Instead, the writing studies community should also look at how computer scoring response to each essay, particularly where human and computer scores do not align, since it is the individual essay-writer who is of prime importance to the field. We should ask, and try to understand, whether automated approaches would handle certain essays better than others. The current frame of reference for AEE foregrounds the proportion of essays that the computer scores correctly (i.e., in agreement with human raters); what has been downplayed is the proportion of essays, even if small, that would be scored differently by computer.

For now, I will simply remark on one commitment that Page's objectors have in common with Turing: both foreground a social conception of intelligence, whether it be the social process of writing or Turing's discursive "test" of intelligence. Writing, seen as an act of grasping and conveying meaning, is a useful and powerful one for many reasons and would seem to reflect Turing's notion of intelligence as a discursive series of acts.

The Legacy of Mimicry

[7] On the view that I have advanced, the "computational implementation" of a framework for semantics that Perelman mentions could be possible on either a Turing-computable or a simulation system.

The bar for Project Essay Grade set by Page—passing a Turing test by mimicking scoring by the use of proxy variables, with the primary success criterion being statistical reliability—has minimized the importance of the existence of a host of other success criteria. It has also seemed to foreclose on other, more normatively defined, questions of how AEE ought to be used.

The next chapter examines how Page's statistical reliability approach has remained the primary criterion by which AEE is judged by those in the educational measurement field. Critics have countered by pointing toward shortcomings of AEE in terms of various forms of validity. The narrative and counter-narrative provided is one of what it means for AEE to properly score an essay.

3. REVIEW OF ASSESSMENT RESEARCH IN AUTOMATED ESSAY EVALUATION

Overview of Chapter

The previous chapter saw that AEE was technologically rooted in the computer science revolution, and symbolically rooted in Alan Turing's indistinguishability test to base legitimacy on human-like patterns of response. But empirical, rather than conceptual, studies of AEE are its lifeblood, where much research has attempted to describe and measure their applications. There is quite a body of research on this technology; indeed, an EBSCO search reveals 101 articles containing the phrase "automated essay scoring" in the past 13 years alone. An annotated bibliography created by Haswell, Donnelly, Hester, O'Neill, and Schendel (2012) provides an "abridged" collection with 39 key sources (para. 3).

Since it would be impossible to exhaustively review all of this research, this chapter will focus on key studies that lay the ground for a new study to be discussed in chapter 4. Along the way, we will see the way in which much of the research reproduced the central argument of Page, that interrater reliability is a form of Turing or indistinguishability test, and that passing such a test legitimizes AEE.

We will also examine challenges to reliability as the appropriate indistinguishability test. Instead, opponents proposed, what ought to be the standard is some form of validity: AEE must be able to evaluate a construct of writing that has relevance or applicability to other writing situations, not simply the construct of writing

created when a software program attempts to reproduce ratings on a highly-managed standardized examination, assigned by people using a highly-constricted rubric. That argument often manifests as some form of humanist objection: humans can capture the full writing construct in their grading, the argument goes, and if AEE is to be legitimate, it ought to be able to capture that construct as well. It should score essays validly as well as reliably. This chapter uses this framing device to see the literature on AEE as a story about what it means for these systems to "get it right."

Current Educational Climate: a Brief Note

With open admissions still representing the entrance philosophy of the vast majority of U.S. colleges and universities, Mina Shaughnessy's (1977) documentation of American college students' struggles with academic writing applies to today's students as well. In short, students need help becoming strong writers. The question of how much automation in educational service delivery is the "right" amount comes in the midst of a major sea change—that of the movement to online course delivery. Massive Open Online Courses, or MOOCs, for instance, reduce classroom contact dramatically and they also provide opportunities for synergy between automation and the virtual course presence typical of the MOOC model. The company responsible for Harvard and MIT's open online courseware, EdX, announced that it would use automated scoring and has developed a tool named the Enhanced AI Scoring Engine (EASE) that is already in use by some of the Massively Open Online Courses (MOOCs) that it coordinates. Christian Schunn's laboratory at the University of Pittsburgh has studied an online peer review platform called SWoRD and the data generated by the huge amounts of coursework it supports, with the hope of developing automated instructional capacity (Schunn, 2013).

(According to Schunn's website, the platform distributes essays, stores evaluations, and even determines the accuracy of the evaluations through some unspecified form of bias identification.)

Not surprisingly, computers score far more rapidly than people, and computer scoring has been deemed robust enough for computers to be employed in a number of standardized tests. The Graduate Management Admissions Test (GMAT) uses a software engine by Vantage Learning called IntelliMetric to score its Analytical Writing Assessment piece in combination with human raters; the Test of English as a Foreign Language (TOEFL) writing section is scored by one human rater and ETS's *e-rater* engine as the second rater with a second human adjudicating if scores differ substantially or the essay is flagged as anomalous (Weigle, 2013, p. 91). The *e-rater* engine is also used to score the writing section of the Graduate Record Examination (GRE) in conjunction with a human rater.

Not to be left behind, both the Partnership for Assessment of Readiness for College and Careers (PARCC) and SMarter Balanced, the two bodies funded by the U.S. Department of Education to develop the assessments for the new Common Core public education standards currently adopted by 46 states—have indicated interest in using automated scoring in the assessments they will develop, following a competition announced by the Hewlett Foundation to incentivize innovation in assessment technologies (U.S. Department of Education, 2012, p. 9).

Although the Educational Testing Service did not provide an exact number of colleges and universities subscribing to its online, automated scoring platform *Criterion*, anecdotal reports from listservs, research documents, and elsewhere point toward a

growing presence of AEE at the college level. Institutions such as Fresno State University, the New Jersey Institute of Technology, Jackson State University, and San Jose State University employ AEE in some form. Outside of the essay scoring realm, automated reading and writing technology is routinely used in human resource departments to scan resumes and will soon be used by the Associated Press to "write" all of its finance reports (Colford, 2014).

In sum, AEE technologies are continually being refined and deployed, over the protestations of many writing experts. Large-scale experimental designs testing the pedagogical impact of AEE in college settings would seem to be the best way to answer the question of whether these applications can improve writing. There are some such studies with mixed results, as will be described shortly. But tracing the evolution of AEE shows that the large scale experimental design was not the first method used to legitimize AEE; that distinction went to reliability-focused correlation analyses examining the scoring patterns between automated and human scoring.

AEE Literature: Reliability Studies

Generally speaking, the reliability of a test refers to the "consistency of scores across replications of a testing procedure" (American Educational Research Association, 2014, p. 33), or in other words the degree to which scores remain the same as a test is re-administered. In the case of essay assessment, rather than the test itself being re-administered, the scoring process is reapplied to the same test samples, by a different rater. Reliability is then calculated as a function of the consistency of the two (or more) sets of scores. This is known as inter-rater reliability.

Reliability is the traditional mode of measuring AEE, beginning with Page and

philosophically and historically rooted, as I have suggested, in Turing's socio-linguistic formulation of a machine-human indistinguishability test. In the case of an AEE system, "reliability" has been taken to mean that the system produces scores that agree to a large and specifiable degree agree with human scores. Albeit a computational achievement, a Turing machine that could fool a human only 30 percent of the time would not be very reliable instrument, nor would an electronic scale that only agreed with a balance 30 percent of the time. If a survey question is written so vaguely that you answer it one way the first time, and another way the next, it would not yield terribly valuable information about the subject of the survey. Reliability is one way of describing how accurate a particular measure is.

As noted earlier, many early studies involved the application of AEE in "high stakes" or important testing situations (e.g., Attali, 2007; Cizek & Page, 2003; Klobucar et al., 2012; Shermis & Burstein, 2003; Shermis & Hamner, 2012). While AEE's potential application is broad, this technology has been primarily studied for its ability to produce scores that agree with those assigned by human raters on standardized, timed examinations with short word counts (e.g., Page, 1966, 1968; Perelman, 2012, 2013; Shermis & Burstein, 2003; Shermis & Hamner, 2012). These studies tended to focus on whether scoring engines are reliable as far as yielding inter-rater agreement when measured directly and with statistical techniques designed to factor out chance and such. (The scores are often also said to achieve intra-rater reliability in the sense that the scoring algorithms will consistently assign the same score to the same essay once the machine is trained.) Several key studies are reviewed below.

In a well-publicized and widely-cited recent study, Shermis and Hamner (2012)

reported agreement results for nine AES platforms, which were trained to score student responses along eight essay-writing prompts taken from six different high-stakes essay writing tests. The authors concluded that "by and large, the scoring engines did a good job of replicating the mean scores for all of the data sets"; most of the mean predictions were within .10 of the (human) resolved score (pp. 21-24). Quadratic weighted kappas (an agreement measure that statistically factors out the role of chance) ranged from .60 to .84 compared to human raters' kappas of .61 to .85 (p. 24).

Despite major press, the scope of Shermis and Hamner's (2012) project was large, probably too large to warrant these authors' sweeping conclusions about AEE. There were, for instance, major variations in the rubrics being used by each data set, their range, how resolved scores were calculated, the average essay length and standard deviation in number of words, and more. Furthermore, the study did not discuss certain factors that may make score prediction easier or harder, besides the range of the scoring scales (obviously, raters using scales with fewer integers have a higher probability of agreement by chance). For instance, the distribution of scores along a rubric could potentially enhance or detract from score prediction accuracy—even if there are six score values on a rubric, if human scores tend toward the mean (do not make substantial use of the lowest and highest scores, for instance) computer predictability could be made easier.

For these and other reasons, Shermis and Hamner's conclusions have been challenged on multiple accounts by Perelman (2013), who noted, for instance, that the study did not include an ANOVA or regression analysis for the machines as a group to see if they performed differently from the human raters on the overall sample (p. 4). (No tests of statistical significance were reported at all, Perelman notes.) Perelman also noted

the "heterogenous mix" of essays that form the basis for the training and test sets, pointing out that four of the eight sets were source-based essay assignments whose rubrics emphasized reading comprehension, not actual writing.

Another objection that Perelman made had to do with the variables being compared. To calculate rater agreement, the two human ratings were compared with one another and an exact agreement percent attained for a sample (in a total of five essays, if two readers agreed on scores for four of these, the exact agreement would be .80). For the computer models, instead of figuring exact agreement with each of the human scores, the variable presented in Shermis and Hamner (2012) is exact agreement with respect to the "resolved score." However, the resolved score provided in many of the samples was not the combination of the two human scores, as is standard practice. Instead, the resolved score was the higher of the two human scores in four of the eight samples, or simply the first reader's score in one other sample. So while the humans are being compared to each other, the machines are being compared to higher of two scores on four of the eight samples. The ability to replicate scores by rounding up, Perelman argued, artificially inflates the machine agreement numbers.

To make his point, Perelman appealed to classical test theory and its positing of a "true score" for each essay, which does not have to be an integer value. If a true score for a particular essay is a 2.8, say, the two human scores are likely to be 2-2, 2-3, 3-2, or 3-3. Any third rater (human or otherwise) who has narrowed the choice to a 2 or a 3 and who knows the resolved score is the higher of the two has a 75 percent chance of getting the score right by selecting the 3. Indeed, Perelman created a table that showed (pp. 6-7) that the scoring engines did better in replicating the resolved score when it was the higher of

the two scores than when it was the sum of these scores, as is the standard way of creating a resolved score.

In another well-known recent study, Klobucar et al. (2012), studied 603 student essays at the New Jersey Institute of Technology, ultimately claiming that the *Criterion* engine is useful for rapid assessment to provide an "early warning system" (p.110) that identifies weak incoming students prior to course enrollment. The focus of this study was to assess whether *Criterion* was a reliable guide—whether its *e-rater* system accurately predicted course grades of the students who used it. The strongest correlation, the authors reported, was actually between human-scored holistic portfolio scores and final course grades (.43); the *Criterion* essay's correlation with course grades was .29.

In the past five years, the domination of large-scale studies by industry researchers has begun to fracture. Tsai Min-hsiu (2012) studied 923 AEE-scored essays from high school students in Taiwan and found that the human-human rater agreement in scoring the essays was much higher than human-machine agreement particularly in the highest and lowest-scored group of essays. (E.g., the Pearson correlation for the two human raters in the lowest, 1-point, group was .929 and the AEE system's Pearson correlation with each human rater was .345 and .385 respectively, where the second AEE data point represents an alternative data set created by assigning a score of "1" to the essays that the AEE system rated unscorable instead of leaving them out of the sample). These results seem to directly contradict industry studies—although Shermis and Hamner (2012) concede that their study did not evaluate potential differential treatment of subgroups by the scoring methodology (p. 27). A separate study of scoring reliability on a 600-word prompt (McCurry, 2010) showed that machine scoring reliability was lower

than human reliability when the writing prompt was more broad and open than the original kinds of prompts found in earlier studies. (Machines had been trained on 187 human-scored essays and produced scores for 63 essays.) McCurry found, for instances that the first software package only awarded scores between 5 and 8 of the nine-point scale, and the second package only awarded scores between 3 and 8. Standard deviations were smaller for the machines than the humans (.79 and 1.18 compared to 1.37). Using fewer points of the score range would be a way of increasing the probability of agreement.

In more general terms, Perelman (2012a) has criticized some of the fundamental reliability findings, proposing that they are only effective for short essays and not relatively longer ones, a claim that McCurry (2010)'s research supports. I discuss this critique in more detail in chapter 4. Ericsson and Haswell (2006) have critiqued AEE on the grounds that most of the studies that show its effectiveness have been done with the collaboration of the AEE industry. These authors have also critiqued AEE from the vantage point of social constructivist pedagogy, contending that writing to non-sentient machines is not a form of writing at all, if writing is indeed a social interaction. This "humanist" objection has been discussed in chapter 1, where I have attempted to show how this objection fits into the larger context of conversations around artificial intelligence from realist and pragmatic philosophies. Herrington and Moran (2001; 2006) have shown that AEE seems unfit to evaluate certain forms of writing, such as essays not written in Standard American English. Still others (Condon, 2013; Perelman, 2012) have suggested that AEE algorithms only measure a partial set of the writing construct, privileging grammar and mechanics over higher-order concerns.

AEE Literature: Validity Studies

The most "fundamental" consideration in designing and evaluation tests, according to the *Standards for Educational and Psychological Testing* (2014), is validity. In the educational measurement context, validity is the degree to which "evidence and theory support the interpretations of test scores for proposed uses of tests" (*Standards for Educational and Psychological Testing*, 2014, p. 11). As some have noted, reliability is a necessary condition for validity: if scores are inconsistent, it is hard to draw conclusions about how they should be used or what they stand for. On the other hand, reliability is not in itself sufficient for validity: measures may be reliable simply because they skim off the more complex qualities of a phenomenon, hence reducing their validity. (E.g., a multiple-choice math test may exhibit high reliability but not test a student's ability to find elegant solutions.)

This relationship between reliability and validity is apparent for AEE as well. If AEE has in its favor reliability in terms of interrater agreement (although as mentioned above, this claim deserves greater scrutiny), the greater challenge is validity: many people question whether AEE scores, as consistent as they may be, are actually measuring any meaningful construct of writing. In particular, Condon (2013) has faulted the assessments for underrepresenting the writing construct in favor of features that AEE engines can score reliably through the proxes they employ. According to Perelman (2012a), "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" (p. 121).

Much of the validity literature for the present project comes from rhetoric and

composition scholars and consists of case study, single institution or single-classroom studies that seek to expose the limitations of grading essays by computer (e.g., Herrington & Moran, 2001; Maddox, 2006; Matzen & Sorensen, 2006). Herrington and Moran (2001) wrote various versions of essays to WritePlacer Plus and Intelligent Essay Assessor (the former is part of the College Board's Accuplacer system and the latter is a feedback engine using latent semantic analysis). The researchers concluded that the scoring methodologies of these two programs were not presently capable of factoring in clichéd language, plagiarism, and other factors that matter in constructing an essay, but focused on measurable features such as essay length, vocabulary, and other simplistic measures of writing ability. Herrington and Moran's main claim was that writing to a machine rather than to a person dramatically shifts the rhetorical situation away from one of creating meaning and impact.

Writing in the Ericsson and Haswell (2006) anthology, Matzen and Sorensen described challenges in using ACT's e-Write software for placement at Utah Valley State College, hoping to find an alternative to the current placement system utilizing multiple choice tests. They reported that the e-Write software had lower correlations with the other forms of placement tests than did the human readers' scores and that, for instance, e-Write would have placed only 4 of 298 students in the lower-level basic writing course, a result that experienced staff deemed too low by historical measures (p. 137). This was an experiment designed to measure validity through convergent evidence. Similarly pessimistic findings on e-Write were reported by Maddox (2006) in a study taking place at Jackson State Community College.

Many industry collaborators have themselves agreed that scoring engines tend to

measure only a "subset" (Klobucar et al., 2012, p. 106) of good writing skills (see also Lim & Kahng, 2012, p. 42). Weigle (2013) investigated several forms of validity in AEE's scoring of 386 English language learners at eight institutions, finding that machine scoring seems to support explanation and generalization forms of validity more than extrapolation, explanation and utility.

Integrating the reliability and the validity evidence gives a mixed view of AEE. It indicates that some measures of reliability (such as aggregate interrater agreement) show computers to be as reliable as people. But examining the validity evidence shows serious doubt that the scores that automated software assigns have much bearing on writing in other contexts, such as college essays or college performance in writing classes.

The reliability story and the validity story provide a distinctly different picture of the merits of AEE. If what it means for AEE to get scoring "right" is to approximate human scoring patterns in highly controlled circumstances, then computer score prediction does have merit. Ellis Page has been vindicated. But if what it means for AEE to get scoring right is to generate scores that provide insight into students' writing ability, aptitude for a particular writing course, or likely outcome of a first-year composition program, then AEE has far to go. There is little reason to expect, 49 years after Ellis Page introduced Project Essay Grade, that AEE models can be used for these more general purposes.

What needs to be done is that we, like Haraway, entertain both perspectives at once. Although the reins driving AEE technology adoption may indeed be guided by the invisible hand of capitalism, AEE may still have underexplored merit. Co-opting AEE for formative rather than summative assessment could be one such avenue, as some suggest.

Formative/Classroom-Based Research

Even as the legitimacy of AEE remains vigorously and acrimoniously contested, Whithaus (2013) noted in his Foreword to *Handbook of Automated Essay Evaluation* that the automated evaluation field is moving toward a wider range of applications than scoring: "the shift [from use of the term 'automated essay scoring' to 'automated essay evaluation'] indicates that feedback, interaction, and an altogether wider range of possibilities for software is being envisioned…" (p. viii). In other words, AEE is not going away and is in fact pushing into other domains. "We expect," wrote Shermis, Burstein, and Bursky (2013) in that same volume, "that there will be greater interest in automated essay evaluation when its use shifts from that of summative evaluation to a more formative role" (p. 10), meaning contexts where the computer program provides some sort of advice to the writer on how to improve writing rather than simply issuing a score. Current research in AEE is growing from studying reliability in standardized testing environments into studying AEE in longer essay assignments, in the classroom as a feedback application, and AEE as an assessment tool for placing incoming college students in ability-appropriate courses (e.g., Isaacs & Molloy, 2010; Matzen & Sorensen, 2006; Klobucar et al., 2012).

However, research on formative feedback engines has been mixed. Roscoe et al. (2011) studied the Writing-Pal automated tutoring system and did find that the revised essays received significantly higher scores than the original ones. However, Chen and Cheng (2008) studied a system of formative feedback called the *MyAccess!* system from IntelliMetric. They implemented the system in three classrooms of advanced English learners and followed up with surveys of students, concluding that, "overall, [its] use was

not perceived very positively by students in the three writing classes" (p. 107). Shermis, Burstein and Bliss (2004) followed 1072 high school students randomly assigned to work on *Criterion* writing prompts or control prompts in advance of a state assessment. No significant difference in performance on the assessment was found in the treatment versus control group.

Hagerman (2011) tested *Criterion* by performing over 60 manipulations to two essays, a *Criterion*-supplied example and a researcher-written one, and submitting them to: a *Criterion* prompt, a scored instructor prompt, and the open-ended "check your writing" module. The manipulations altered features thought to contribute to *Criterion*'s scoring algorithm. Hagerman, a computer scientist with experience as an ESL instructor, concluded that the "single greatest determiner of score is simply the total word count" (p. 281). He further concluded that "the results do not support the claim by ETS researchers [in Burstein, 2009] that *Criterion* 'embodies' Process Writing except in the most superficial way" (p. 281). As Lim and Kahng report, Cheville (2012) found that *Criterion* may discourage "writers' development and use of logic and imagination" (p. 44).

AEE Enters the Classroom

As indicated, AEE has been conceptualized by Page and others as a tool to generate reliable scores, in which aggregate score assignments appear to be on par with those of human raters. This dissertation offers additional framing language designed to offer insights into how this technology serves or does not serve students as individuals and learners. In what follows, that new framework will be narrowed even further to college students.

A small study (described in the next chapter) will explore whether computer

scores align with how college instructors might score papers. The warrant for asking this question is that if computer scoring and feedback is to be useful to college writers, it would help if the software is able to assess essays in a way consistent with the kinds of grading and feedback students might expect from their instructors. Seeing how well AEE scores match up with trained instructors' scores is a way of assessing this kind of validity.

If scores do not align closely, this would suggest that AEE not be used as a general scoring mechanism. It may yet be valuable for students to write practice papers to automated assessment software in which they receive grammar and mechanics feedback, but AEE systems should in that case not promote nor leave the impression that its scores measure or correlate with anything more than the features that make up its feature set.

Thus, for the essay scores generated by AEE to be relevant in higher education, according to my premise, they ought not only be reliable on what Deane et al. (2013) has called a "common rubric" that measures surface features and structure, but their scores ought to cohere with those along a rubric that operationalizes the features that college instructors tend to value, such as those in the Framework for Success. While Attali (2013) did report findings from some studies on the relationship of automated scores with independent measures of writing ability such as self-evaluations, far more work is needed to characterize the relationship between automated evaluation drawing on surface features, and rubrics suitable for the expectations for college writers. In order for the machines to be useful in college settings, they must be able to evaluate papers in ways similar to how teachers might evaluate them. This intuition needs to be operationalized, and I will discuss the operationalization in what follows.

Some qualities thought to support the kind of writing tasks navigated in college

are those enumerated in the Framework for Success in Postsecondary Writing and the related Outcomes Statement. The Outcomes Statement is an attempt to characterize the "knowledge, practices, and attitudes" that it is hoped students will acquire through first-year composition courses (CWPA, p. 1). These outcomes include rhetorical knowledge, critical thinking, writing processes, knowledge of conventions and ability to compose in multiple environments—although the last outcome has been removed in the 2014 version in the recognition that multimodal composing is implied in the four previous outcomes (O'Neill, Adler-Kassner, Fleischer, & Hall, 2012). (See Figure 2 for list and abbreviated descriptions of Outcomes.)

CWPA Outcomes

- Rhetorical knowledge – the ability to analyze and act on understandings of audiences, purposes, and contexts in creating and comprehending texts;
- Critical thinking – the ability to analyze, synthesize, interpret, and evaluate ideas, information, situations, and texts;
- Writing processes – multiple strategies to conceptualize, develop, and finalize projects;
- Knowledge of conventions – the formal rules and informal guidelines that define genres, and in so doing, shape readers' and writers' perceptions of correctness or appropriateness

Figure 2. List of Framework Outcomes (2014).

Perelman has asserted (2012a) that AEE would not be able to measure most of these abilities. He alleges that *e-rater*, and probably IntelliMetric and Intelligent Essay Assessor as well (although these software developers have been less transparent about their scoring methods) primarily count errors, words, and characters (p. 129).

This is a good place to raise an issue that has been found in AEE scoring, which is that, as noted by Shermis et al. (2013), a correlation has been found between the length of an essay in number of words and its human-assigned score. After a certain range of lengths the correlation begins to recede (pp. 8-9) and some would put that threshold at about 1,000 words. Accordingly, the length-score association may well tail off for college-length essay assignments, which may range to 1,500 words for a typical first-year composition assignment. Furthermore, timed writing assignments produce essays of dramatically variant lengths (e.g., the same prompt may result in essay of 100 to 600 words or in other words, a longest essay six times as large as the shortest one). In contrast, a typical range for a college essay might be that the longest sample is twice as long as the shortest (e.g., 5 pages to 2.5 pages). Assuming that scores are normally distributed, essay assignments where length is a more prominent, noticeable feature (the length multiple between shortest and highest is larger) would seem on the surface to stand a better chance of having a higher length/score correlation.

To continue with this example, even if a length-quality association were found to hold in college writing, certainly it would not be because length alone is intrinsically valuable. Consequently, if a computer system were to provide meaningful feedback in addition to a score it could not do so by advising the student to just write a longer paper. To wit, as Burstein, Chodorow, and Leacock (2004) noted in a white paper introducing

ETS's *Criterion* Online Writing service, "…[I]t is essential that students receive accurate feedback from the system with regard to errors, comments on undesirable style, and information about discourse elements and organization of the essay. *If the feedback is to help students improve their writing skills, then it should be similar to what an instructor's comments might be*" (p. 4, italics mine). In formative feedback, score estimation by proxy variable is not sufficient unless the computer is able to capture the student's performance like a teacher would, which would only be possible if the variables measured are identical to or closely-enough related to the variables that do matter to college instructors.

Moreover, according to ETS researchers Chodorow and Burstein (2004), *e-rater* contains no "direct measure" of an essay's length. The cited reason is that the "driving concept that underlies *e-rater* is that it needs to evaluate the same kinds of features that human readers do….Even though length measures can be shown to be highly correlated with human reader essay scores, length variables are not scoring guide criteria (Burstein & Chodorow, 1999, p. 69, cited in Chodorow & Burstein, 2004, p. 4). Chodorow and Burstein (2004) go on to remark that "Although some researchers have suggested adding explicit measures of length to improve correlations with human readers, the assessment community is generally concerned about the effect this would have on coachability" (2004, p. 4). But, the authors do acknowledge that several features of e-rater could serve as "proxies" for essay length in the models that the system builds (p. 11).

Returning to the Framework and its Outcomes Statement, ETS researcher Deane (2013) also recognized that the kinds of values embedded in the Framework are not measured directly by *e-rater*, in contrast with the values found in more traditional holistic rubrics:

> These systems are designed to be applied in the first instance to holistically scored, on-demand, timed assessments. In general, AES systems rely on features intended to measure many of the traits specified in holistic scoring rubrics, or in the Six-Trait model offered by Spandel (2005). The systems match much less well with criteria advanced in such contexts as the Framework for Success in Postsecondary Education (Council of Writing Program Administrators, National Council of Teachers of English, and the National Writing Project, 2011; O'Neill, Alder-Kassner, Fleischer, & Hall, 2012), which emphasizes rhetorical knowledge, critical thinking, and control of the writing process. In part this disjuncture is due to the contrast between models focused on "text quality," measured in the end product, versus models focused on "writing skill," which is an attribute of the writer, not the text. In part, it reflects differences in what kinds of features can readily be measured in the current state of natural language processing technology. (p. 12)

As indicated above, current natural language processing technology does not readily measure the kinds of qualities captured in the Framework. In fact, all parties seem to agree that, as Deane et al. (2013) wrote, "as the articles collected in Shermis & Burstein (2003) indicate, AEE systems do not explicitly evaluate the validity of reasoning, the strength of evidence, or the accuracy of information. They rely instead on measures of such things as the structure and elaboration of student essays, the

sophistication of vocabulary, or the number of errors in grammar, usage, mechanics or style" (sect. 1). Ultimately, many would agree with Klobucar et al.'s conclusion that, "writing program administrators take it as a given that only the course instructor can witness robust construct representation and evaluate it by awarding a course grade" (2012, p. 79).

Criterion's feedback is a function of the traits it is able to measure. *E-rater's* holistic score is calculated based on traits or features of writing, which as Quinlan et al. (2009) noted, blurs the traditional distinction in assessment between holistic and trait-level scoring. Because it cannot comment on features it cannot score, *Criterion*'s ability to value essays comparably to how college instructors value them, and therefore to provide meaningful feedback, depends on how closely its scoring matches up with how instructors would score. (If it can't measure what matters, it won't instruct properly.)

The study to be described in Chapter 4 poses questions about the inter-correlation between *Criterion* holistic scores and human scores along a rubric that prioritizes higher-order skills such as those found in the Framework. As Paul Deane noted in both Deane et al. (2013) and Deane (2013), stronger writers tend to be able to master many of the components of good writing, whereas weaker ones may have multiple deficits. This "deep connection between skill in text production and the broader bundle of skills deployed by expert writers" may be accounted for by the theory of cognitive capacity, whereby writers who have achieved fluency in writing have more cognitive resources available to them to tackle higher-order dimensions of writing (Deane, 2013, pp. 17-18).

Synthesizing the remarks made above, it turns out, Paul Deane has provided both a rationale for why AEE would *not* be able to evaluate essays on qualities embodied in

the Framework's Outcomes Statement, but also a rationale for why AEE *might* be able to. While AEE does not directly measure aspects of writing such as quality of an argument, it might indirectly do so because of the cognitive relationship between higher and lower order features of writing. If AEE systems can exploit writers' cognitive capacity they may be able to perform reasonably well on writing assignments judged with Outcomes qualities. An interesting research question, then, could be framed around the extent to which there is an overlap between common writing qualities and higher-order Outcomes sort of qualities. This is one motivation for the pilot study described in chapter 4.

How AEE Works: Notes on Mechanism

A functional understanding of automated scoring will help structure the discussion in chapter 4. I concentrate here on one of the most popular scoring engines, *e-rater* from the Educational Testing Service, which is the scoring engine that was used in the study described in that chapter.

While the features that *e-rater* uses change periodically, the current version of ETS's *e-rater* engine (version 11.1) lists the high-level features of:

- Grammar
- Usage
- Mechanics
- Style
- Organization
- Development
- Positive features

- Lexical complexity
- Topic-specific vocabulary usage

Each high-level feature is composed of microfeatures that *e-rater* measures and figures in to calculating the final scores, through a regression analysis. (A regression analysis is a way of predictively modeling a phenomenon by manipulating a set of variables.) See Figure 3 for the construct decomposition or a conceptual map of the features deployed in the latest version of *e-rater*, including the list of microfeatures that serve as variables in the regression model (reprinted with permission from Deane, 2013).

As Deane (2013) noted (p. 14), high-level features are grouped into nine headings that correspond to commonly-used rubric categories in writing assessments. The six traits in the box on the right hand side are from the 6+1 rubric (Coe, Hanita, Nishioka, & Smiley, 2011). The 6+1 rubric grew out of assessment research that Paul Diederich and others conducted at ETS. In his (1996), Diederich conducted a factor analysis seeking to elicit the writing qualities driving expert grades in a sample of 300 freshmen essays. For his analysis, teacher comments were tabulated and grouped into the dimensions of *ideas, mechanics, organization, wording and flavor* (Diederich, 1996, p. 358).[8]

[8] In a study of the comparative validities of four scoring systems, Winters (1980) found that the Diederich scale correlated mildly with both impressionistic and analytic scoring systems and was able to distinguish between high and low-performance writers, an indication of its predictive validity.

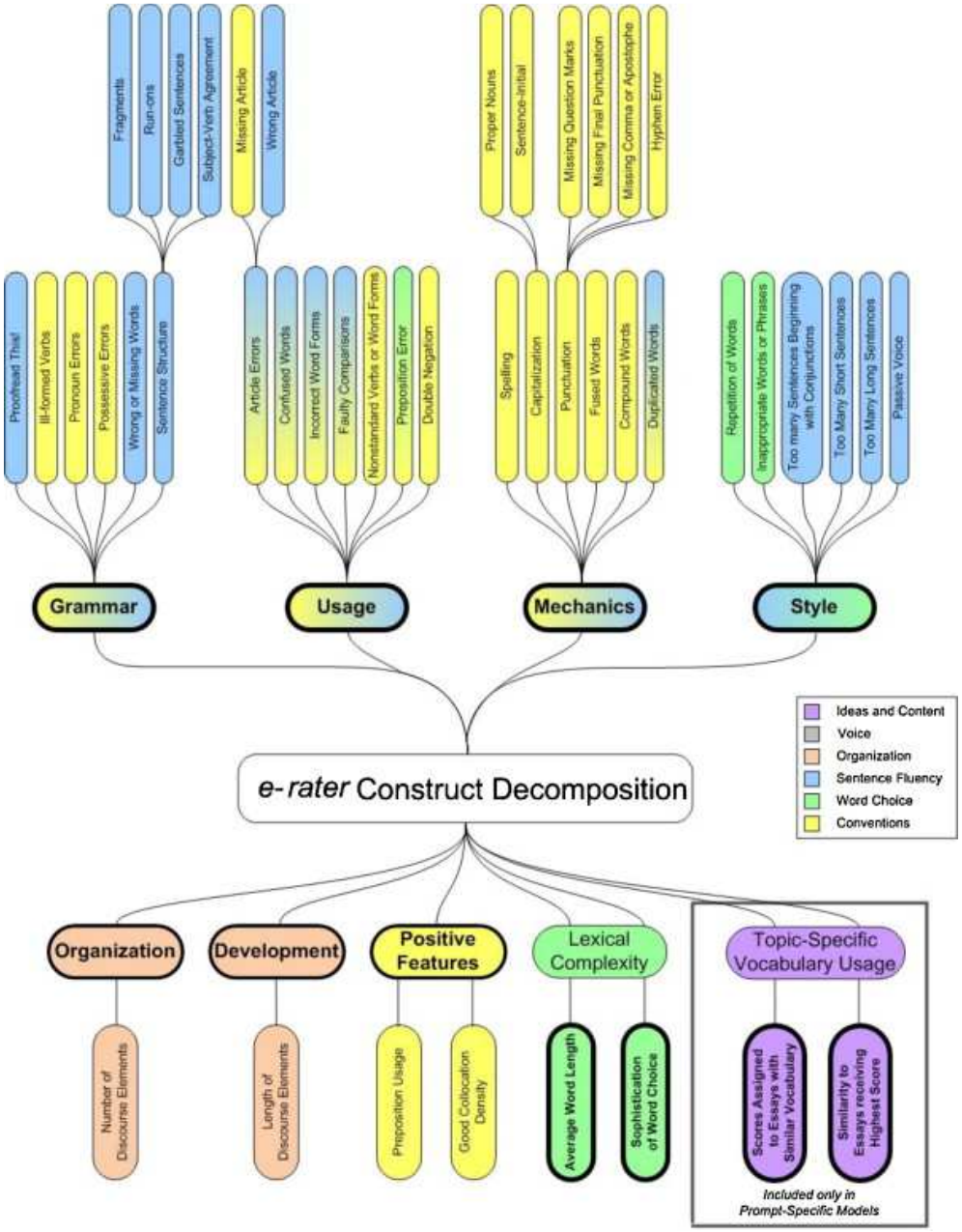


Figure 3. Construct decomposition for *e-rater*.

The visual presentation of the construct decomposition is of course the result of rhetorical decisions, not simply computer engineering constraints. There was a decision to present *e-rater*'s high-level features as the same gestalt features that people would

employ to react to an essay. In other words, it presents an image of the computer as a reader, judging style or organization as a human would. In fact, though, as indicated by the blurring of the colors in each feature, corresponding to the key at the right, the microfeatures that the computer algorithm actually counts could be arranged in different ways; for instance, Article Errors under "Usage" is actually said to correspond both to Sentence Fluency and to Conventions. Even more, the high level features themselves do not presumably contribute to the score; rather, the microfeatures do. While "grammar" and "style" typically represent meaningful and causative categories for a human rater, they do not represent meaningful nor causative categories for *e-rater*, as it is only the microfeatures that contribute to a score.

Like the appeals to Turing that Page made 49 years earlier, this kind of a visual and textual presentation of the scoring features is a part of the symbolic engineering of AEE technologies, emphasizing how the computer compares (favorably) to humans. Those opposing automated scoring have labored to present computers and people as antitheses, such that the sponsors of the petition against automated scoring have named their website "humanreaders.org," to highlight the *prima facie* absurdity in the premise of machine reading. In contrast to the rhetorical message presented in the construct decomposition chart, the study of computationalism in the preceding chapters uncovered that computers are in this case simulating not replicating human scoring.

As many have observed, the microfeatures that *e-rater* records do not always relate to features that readers themselves would take into account or, in some cases, even be able to identify. While human judges may certainly take spelling into account, and some of the grammatical microfeatures such as subject-verb agreement, they would never

precisely calculate an essay's average word length or length of discourse elements such as paragraphs and factor those into their score. Most instructors have thought little about "good collocation density" yet manage to grade student papers. Now one might argue that such features may operate in the background as the linguistic correlates of what raters might squishily refer to as "good style." But these relationships have yet to be identified—although C. Anson (personal communication, 2014) reports that in one study, a higher proportion of Latinate versus Anglo Saxon vocabulary resulted in perceptions of higher quality. What is obvious is that these *e-rater* microfeatures are not found among the formal characteristics of evaluation, not found on any rubric nor discussed in any scoring session. Chodorow and Burstein's (2004) claim that computers evaluate features similar to what humans evaluate is not quite right: they simulate evaluating the features that humans evaluate—or at least are asked to evaluate via rubrics like 6+1.

Indeed, the microfeatures in Deane's analysis correspond to what Ellis Page would have called "proxes" or approximations of intrinsic variables of interest in measuring the quality of an essay (discussed here in chapters 1 and 2). The macro-features correspond to intrinsic variables of interest or "trins." Dyehouse (2007) suggests that knowledge had been so consolidated around the point that computers only simulate a true analysis of an essay such that by the 1990s it was not necessary to justify the use of proxes, because everyone knew that computers did not actually read as a human would. However, what Deane and predecessors at ETS seem to have been able to do is to fuse the master analysis and the ratings simulation in the visual representation. In this representation, the proxy measures at the microlevel contribute to the macrolevel master analysis.

In sum, as many have pointed out, automated scoring engines work with text in ways that are quite distinct from how readers work with text, manipulating different features. The scoring process is very much, then, a simulation rather than a replication of the human undertaking. And the visual representation as given in Figure 3 does not mimetically represent the computational calculus—the computer's methodology—as much as it *symbolically engineers that calculus* so that it corresponds to the human methodology for scoring.

This mode of similarity-building mirrors the standard measure of success in automated scoring that the computer scores should be able to predict the average of two human scores on a set of essays and that the computer scores attain similar levels of agreement as the two human raters (c.f. Deane, Williams, Weng, & Trapani, 2013). As pointed out in chapter 2, the idea that computers should be measured in terms of their ability to agree with or mimic human raters of course reflects Alan Turing's (1950) proposal that machine intelligence can be measured by a computer's ability to pattern its responses after a human respondent so effectively that it fools a referee (now called the "Turing test"). It also stems from the idea that there is no other agreed-upon "gold standard" for evaluating essays as there would be in, say, arithmetic performance where the answer to a straightforward multiplication problem such as 25 x 59 is arrived at through an agreed-upon method.

The verdict that automated scoring is a "fait accompli" (Cizek & Page, 2003) is premature and obscures vigorously contested space. The jousting match has begun to settle in the statistical details and psychometric methodologies of the studies, as we saw above in the dissensus between Shermis & Hamner (2012) and Perelman (2013). A meta-

analysis of scoring studies, with attention to methodological considerations such as how interrater agreement is calculated, would surely yield insight into the reliability of scoring engines and how that reliability is affected by choices in study design.

The study carried out in this dissertation, described in the next chapter, demonstrates how a subtle shift in perspective from how closely machine and human scoring correlates on average, to where and how the scoring differs, illuminates the impact of high-stakes and machine scoring on students. In order to assess the potential value of *e-rater* for college writing, the next chapter explores what conclusions can be drawn about the relationship between its scoring approach and the approach of a presumptive college instructor. This can be done across a range of essay types: the default essays that *e-rater* has pre-adjusted scores for, and essays assigned by instructors. The study described in chapter 4 uses the first kind of essays.

While some have pushed AEE to absurd extremes by feeding it meaningless but grammatical essays and showing that the programs do not recognize gibberish as such (as Les Perelman has done), or by feeding it essays written by famous novelists and showing that the machines do not recognize excellence (Wilson 2006), I am less interested in "stumping" the machines where we know they do poorly. I am interested in studying the machines where they would be likely to be used, such as by assigning them to score short-answer responses that might be used to place students in college courses or to quickly assess their general writing skills. In looking at score divergence my hope is to uncover some quantitative and qualitative insights into what automated scoring means for students.

Why a Rubric? A Response to Skeptics

Not only has the writing assessment community disagreed over the best measures for the success of automated writing evaluation—whether statistically reliable scoring or valid assessments—many of its members reject the notion of reliability more broadly.

In assessment, reliability of an instrument—the consistency of its scores across multiple dimensions—is a fundamental requirement. Why, then, asked Ed White, "does the Americanist on the assessment committee quote Ralph Waldo Emerson, 'A foolish consistency is the hobgoblin of little minds?'" (White, 2012, p. 500). Here, White pointed out a certain "gulf" between the college composition and educational measurement communities about rigorous assessment (p. 499), concluding that it represents a conflict over values. In the arts, people naturally value disagreement (over Cassatt's best painting or Woolf's strongest essay, say) but in assessment, a researcher has to be able to show that a scoring instrument can be applied by different raters to produce similar results. If a student essay is both a work of art and a data point, the conflict commences. It is a conflict where, White insisted, espousing a pluralist position, "nobody is 'wrong' in this diversity of world views" (p. 501).

Because my study utilizes an assessment methodology that presumes that reliability is an important indicator of the stability of the concepts of writing ability being measured, we should examine two versions of objections to assessments and rubrics. The objections I will consider fall under these categories: (1) appeals to critical theory or qualitative ideology to voice suspicions of quantitative assessment (e.g., Scott & Brannon, 2013; Broad, 2003, 2012) and (2) concerns about the application of global or unitary rubrics to local circumstances (e.g., Anson, Dannels, Flash, & Housley Gaffney, 2012).

The existence of this contested ground of assessment was recently exemplified in an essay by Scott and Brannon (2013), who in their qualitative study and critique of their university's writing program assessment sought to "remake the assessment scene, responding to Chris Gallagher's call in CCC to take assessment out of the exclusive hands of administrators and the testing industry and to assert the agency of teachers and students who are 'there'…" (p. 275). Here and elsewhere, assessment is critiqued as a tool imposed by those in power (administrators and industries) to silence dissent and solidify their privilege. In this view, assessment is a site of "struggle for the democratization of education." The authors cast their critique as distinct from the classical divide between the educational measurement and college composition ideologies, because "while it has been well argued in writing assessment scholarship that assessment should take into account different value judgments about what constitutes 'good writing,' how assessment relates to labor structure and practices in first-year writing programs has not been extensively explored" (Scott & Brannon, 2013, p. 275). To correct the record, these authors aim to provide a social critique of the material conditions underlying assessment.

As White has implied, assessment critics often target the quantitative methods behind the practice. Quantitative assessment entails the development of common standards that ensure inter-rater reliability. Such standards, argued Scott and Brannon (2013), do not arise as the outcome of a harmonious consensus but misrepresent the messy reality behind which voices are effaced and which are heard, which in turn devolves around power relationships. The contingent faculty Scott and Brannon studied held different "chronotopes" of literacy from the higher-status tenure-line faculty, the

former attending more to formal characteristics of text, the latter inclined toward dissensus and heterogeneity. Moving toward a consensus could conceivably mean privileging one set of values over another. As these authors argued, "programmatic writing assessments that lack qualitative elements remain pervasive in no small part because these assessments align well with exploitive labor practices, shifting the focus toward measurable characteristics of texts and away from the terms of labor that produced them" (p. 275).

Scott and Brannon's concern follows a familiar trope in composition studies and other areas of the humanities, which sees the "scientific method" as aiding the material interests of the current social order. James Berlin's (1987) study of current-traditional rhetoric located this tradition as embracing the idea that the scientific method could decide questions of value as well as clear-cut questions of science. Such a misplaced belief, Berlin argued, ended up reinforcing the status of the emerging college-bound meritocracy (p. 36-37). Without offering sustained speculation as to the cause, Haswell (2005) documented that the National Council of Teachers of English (NCTE) and Conference on College Composition and Communication (CCCC) increasingly de-emphasized "RAD" (replicable, aggregable, data-supported) research over the past 20 years.

For Foucault (1975), the emergence of the natural sciences traces its origins to the investigatory practices characteristic of the Inquisition. The emergence of the human sciences originated from a related sort of practice of disciplinary analysis. Foucault held that disciplinary analysis never left the human sciences, but remained in the form of school psychology, testing, and so forth.

> For, although it is true that, in becoming a technique for the empirical sciences, the investigation has detached itself from the inquisitorial procedure, in which it was historically rooted, the examination has remained extremely close to the disciplinary power that shaped it. It has always been and still is an intrinsic element of the disciplines. Of course it seems to have undergone a speculative purification by integrating itself with such sciences as psychology and psychiatry. And, in effect, its appearance in the form of tests, interviews, interrogations and consultations is apparently in order to rectify the mechanisms of discipline: educational psychology is supposed to correct the rigours of the school, just as the medicinal or psychiatric interview is supposed to rectify the effects of the discipline of work. (p. 308)

A Foucaultian analysis of AEE would no doubt be fruitful: his characterization of the examination as a form of social control—a way to hierarchically observe and normatively judge—invites analysis with regard to automated (and non-automated) standardized testing. Narrating the history of AEE against a broader genealogy of computationalism would also be compelling. The idea that early field builders such as Page, relying on Turing and others, not only provided evidence for AEE but actually determined what counts as evidence (e.g., simulating human graders) fits well with the ways in which Foucault articulated the evolution of knowledge as created in the context of powerful interests and by power itself. That AEE is being pushed by well-funded

testing companies such as ETS, funded by influential foundations such as Hewlett, and adopted by the national common core standard-setting organizations PARCC and SMARTER-BALANCE only reinforces need for critical analyses of AEE.

That said, Haraway's mandate was to entertain both the fearful perspective about technological advance and the hopeful one. While the investigation may be connected with some form of authority, and while that authority may quite often be the kind of elite authority that troubles Berlin and Foucault (e.g., the state, the bourgeois meritocracy, global capital interests), it is quite a different thing entirely to argue that *all* instances of scientific investigation are *necessarily* servants of either state authority or global capitalism. (Possibly *p* does not entail necessarily *p*.) In other words, which power/knowledge relationships align is a contingent matter.

In the case of assessment, the consensus-building process of assessment may bring strange bedfellows into agreement with one another. Scott and Brannon (2013) did not describe, apparently, the process by which they developed two distinct rubrics to illustrate their point that different values yield different measures and ways in which student work "sort[s] out differently" (p. 290). (One of their rubrics favored formal qualities of writing and the other expressive qualities.) A plausible scenario might be that the administrators who called for the assessment would identify more closely with a formalist rubric, as it is certainly closer to what the layperson may understand to be "good writing." But this would align the administrators with the contingent faculty, rather than the (higher-status) tenure-line faculty. Not that such an ideological alignment would necessarily (or even likely) improve the material conditions of contingent faculty, but that

assessment methodologies, including those of consensus-building—do not always align to support some sort of master authority.

In fact, the essay assessment project I describe in chapter 4 seeks precisely to study the relationship between a "received" and formalist rubric (the rubric that guides the performance of *e-rater*, with the features presented earlier in Figure 3) for measuring writing in standardized settings and a dissenting rubric that derives from values heralded within the composition community. While participating instructors in a study such as the one to be described try to develop a common understanding of a specific rubric through norming sessions and the like, which sometimes does involve acquiescence to a presumed authority, there are times in this norming process when dissensus results in the opposite outcome. For instance, a score on a norming essay may be altered after input from the dissenting instructors. Either way, the norming process in my study will lead to the development of a common understanding of a rubric *other than* one comprising the formalist features favored by AEE and the testing industry, and troubling to compositionists. So again, while we can accept Scott and Brannon's observation that consensus-building is sometimes a messy struggle that involves power at the micro-level (a point to which I and probably many other assessment committee members can attest) we can cleave that point from the broader worry that such consensus-building as a part of the scientific method *necessarily* serves the societally powerful at the macro-level. A more nuanced understanding of power should be put in place.

Another objector to rubric-based assessment, Bob Broad (2003), also held that assessment methods impose an unnatural (and unscientific) sense of agreement, leaving out generous amounts of disagreement over what matters in assessing writing. Broad

(2003) wrote that ETS researchers, among the founders of modern writing assessment, Paul Diederich, John French, and Sydell Carlton, "traded in the rhetorical truth confronting them (that readers value texts differently) in exchange for the grail of high inter-rater agreement" (p. 8). Broad alleged that their use of factor analysis and other techniques to reach common scoring criteria, laboring under the experimental paradigm, only took on the "appearance" of objective truth by "turning away from the messy facts at hand" (p. 7). For Broad, the move to cull—and then impose—common scoring criteria simply was poor science.

Consequently, in his own research on assessment practices at a public research university, Broad developed a qualitative form of inquiry he called "dynamic criteria mapping," which avails itself of the research methods of "constructivist" grounded theory to analyze observational notes, transcripts, and program documents to develop a map of an institution's "terrain of rhetorical values" (p. 24). The end result of Broad's inquiry was not a rubric but a conceptual map that positioned 46 textual criteria, 22 contextual criteria and 21 other factors that Broad assorted in a conceptual grouping with one another, represented in a visual map (p. 21).

For Broad, rubrics are an impediment to knowledge; they are "traditionally the main obstacle to telling the full and true story of how writing is valued" (p. 122). As one of the pitfalls of rubrics, Broad cites that they generally omit contextual factors of scoring—features such as program expectations, student background, and so forth. "Rarely do scoring guides venture into the realm of evaluative context when investigating or reporting on how rhetorical judgments are made," he writes (p. 73).

Rubrics may not, as Broad suspects, capture all of what is harbored in the minds of diverse groups of raters. And the values they do capture may be those of whatever power structure is in operation, as Scott and Brannon imply. However, the purpose of the present study is not to create an ecumenical rubric that either (a) gives a comprehensive picture of the rhetorical values of any given body of readers, or (b) fully represents the construct of writing, which is, after all, an abstract and contested concept. (After the mid-twentieth century fall of conceptual analysis in philosophy of language, many scholars would agree that precisely defining a concept such as "good writing" or "happiness" is an incoherent endeavor, anyway.) Instead, the purpose of my study is to examine the performance of a rubric encoded in an automated software program, against the performance of a rubric that could be said to represent at least some of the shared values of the composition community, and to explore the framing of such an examination itself. Because the goals of the study differ from the kinds of assessment projects that trouble writers like Broad and Scott and Brannon, the use of a rubric (rather than dynamic criteria map or critical inquiry) is appropriate here.

An additional point about the importance of supporting this kind of replicable inquiry, using a specific rubric, is stated by Haswell (2005), who mourns the disappearance of RAD inquiry in the field:

> Even now, the profession's immune system—its ability to deflect outside criticism with solid and ever-strengthening data—is on shaky pins. It lacks a "systematically produced knowledge" (Carr & Kemmis, 1986, p. 8) to defend its central practices from outside attack, lacks a coherent body of

> testable knowledge connected to class size, computer pedagogy, group work, part-time teaching, interdisciplinary instruction, 1st-year sequenced syllabi, and the list can go on. And in part, it does not have the body of facts because its most prominent professional organizations, NCTE and CCCC, do not valorize or support the apparatus needed to drive RAD research. (p. 219)

Much of the RAD inquiry into the merits of automated essay scoring is being done by testing companies. Adopting a discrete rubric is needed in order to address the discourse communities already studying AEE.

A more philosophical objection to assessment, which can only be mentioned here, derives what I am going to call the psychology of evaluation. It contends that it may not be psychologically possible to shunt aside differential preferences just by enough norming on a rubric, because each person's evaluation depends on his or her individual conceptual map. Perspectives from cognitive science suggest that evaluators construct a reading of a student text based on contextual and textual cues. Kintsch (1998) proposed the "situational model" referred to in Mason et al. (2013) in their study of fourth grade readers, which subsequently came known as the construction-integration (CI) model that proposes that meaning is not simply extracted from text, but constructed through a process that activates a reader's memory and those aspects of memory that represent that concept. As a very simple example, the meaning of the word "tree" may activate my recall of concepts such as pine trees, palm trees, or climbing a tree, and that recall is fairly stable for each individual but not fixed, as slightly different memories can be

activated each time. Comprehension is first constructed through visual processing of a text and then integrated as context effects and prior knowledge become activated.

As Schwegler (1991) noted, those who evaluate writing are of course themselves readers, who are likewise engaged in a multi-stage cognitive process in which the mind "arrives at an initial estimate of the text's purposes and design" before attempting "to build a coherent mental representation of a text's meaning" (p. 215). Since these representations involve preexisting memories and other mental states, not only textual features, reading is an individually-determined process; ultimately, so is grading. This concern about the validity of rubric evaluation is going to be postponed until chapter 5, which examines construct-irrelevant variance in human scoring.

Which Rubric?

Because the study described in chapter 4 studies the validity of *e-rater* as a model scoring engine, the choice of a rubric against which to measure *e-rater*'s performance matters greatly. One question is whether general rubrics are acceptable, or whether only those created for particular assessment contexts should be used.

Many have wrangled over the relative merits of global or general versus local rubrics. Moxley (2013) wrote about the value of a general rubric at the University of South Florida, which has used the same rubric to assess over 100,000 student essays since 2011. Moxley views the organizing effect that rubrics have on readers, in the sense of enabling them to achieve some version of consensus, as positive. Moxley described the scoring guide as a "community" rubric used across the writing curriculum at the University of South Florida. It is an analytic rubric, used from 2011-2013, that includes the five attributes of Focus (Basics and Critical Thinking), Evidence (Critical Thinking),

Organization (Basics and Critical Thinking), Style (Basics and Critical Thinking), and Format (Basics). Basics and Critical Thinking refer to two separate traits by which essays were evaluated. Perhaps more consciously sensitive to the variability of the writing task at hand, Pagano, Bernhardt, Reynolds, Williams and McCurrie (2008) described an inter-institutional rubric that consisted of Task Responsiveness, Engagement with Text(s), Development, Organization, and Control of Language.

However, as Moxley himself noted, several key reviewers in the area of assessment have disdained the use of general rubrics to evaluate multiple genres of writing, in favor of locally constructed rubrics (e.g., Anson, Dannels, Flash, and Housley Gaffney, 2012). As those authors concluded, "generic, all-purpose criteria for evaluating writing and oral communication fail to reflect the linguistic, rhetorical, relational, and contextual characteristics of specific kinds of writing or speaking that we find in higher education" (Anson, Dannels, Flash & Housley Gaffney, 2012, "Going Local" para. 1).

Because composition studies tends to favor locally-constructed rubrics, it not surprising to those within composition studies that the Framework Outcomes Statement has not been crafted into a globally-used, standardized rubric. Indeed, the Outcomes Statement language specifies that its focus is upon results and not "standards" or levels of achievement (p. 1). One might think that if these measures are put forth as outcomes that college writing values, these outcomes would have been turned into a measurement tool by which college writing could be assessed. But since it has not, any project making use of the Framework outcomes will not have the advantage of having a rubric that has been shown to produce reliable scores. Any writing evaluation study based on this Framework stands to be as much a study of the Framework as the essays themselves. The next

chapter describes the adaptation of Framework skills into a holistic rubric used in order to evaluate a set of student-written timed impromptu essays.

Use of a Cross-Correlation Methodology

The method of applying different rubrics or scoring valuations to the same prompt or writing assignment has been employed by, among others, Follman and Anderson (1967), Winters (1980), and Deane, Williams, Weng, and Trapani (2013). Winters, noting the shifting meanings of socially constructed concepts such as writing competence, observed that "the method selected for judging writing samples *ipso facto* defines 'good writing.'" (1980, p. 1). The choices made in creating and applying an assessment create a construction of what makes good paper. While such constructions may indeed be derived from communal understandings of writing, they are certainly not fixed across communities (and perhaps not within them either), let alone do these constructions serve as representations of Platonic ideals. And most importantly, as Winters noted, such constructions *construct* writers themselves in quite material ways, because "decisions about writing competence derived from one scoring system may not lead to the same decisions about an examinee when another scoring system is used" (1980, p.1). Winters, then, set about studying the comparative reliabilities and validities of four distinct scoring systems—to understand how the systems corresponded to one another and how they would have sorted writers into pre-established *criterion* groups.

As Winters (1980) pointed out, Follman and Anderson (1967) had previously performed a cross-validation of five scoring systems and found that four of the five systems they examined intercorrelated highly (Follman & Anderson, 1967, p. 198). Their study had ten "themes" or essays evaluated along five scoring systems by five raters

assigned to each system. Correlations between three of the five scales ranged from .935 to .99. Winters (1980) also found high intercorrelations (Pearson *r*) between three of the four scoring systems in her study, ranging from .79 to .86 (p. 26). Interestingly, Follman and Anderson found that the Diederich Expository Scale (DES), or what was then called the Diederich Rating Scale, failed to inter-correlate as highly as the other rating systems they used (correlations ranged from .59 to .61), whereas in Winters' study, the DES performed as well as the other scales, correlating at .82 and .86 with the other two scales that successfully predicted the criterion groups (1980, p. 10-11).

One important lesson learned from these earlier studies—particularly from what happened to the Diederich scale, is that the degree of correlation of any given scale is going to be relative to the set of scales under consideration. One of the findings in the AEE literature, noted earlier in this chapter, is that many automated systems can perform to a certain standard of reliability when raters are using the sort of scale that measures mechanical and organizational qualities of writing. Indeed, Deane et al. (2013) spoke of a "division of labor between human and automated scoring" because "current-generation AEE engines provide direct evidence only with respect to some aspects of writing skill, typically involving features that measure such things as basic essay structure, vocabulary, style, grammar, usage, mechanics, and spelling" (sect. 2). On the issue of addressing higher order writing features such as rhetorical skill and argumentative quality, Deane et al. (2013) the articles in *Automated Essay Scoring: A Cross-Disciplinary Perspective* (Shermis & Burstein, 2003) to claim that "AES systems do not explicitly evaluate the validity of reasoning, the strength of evidence, or the accuracy of information. They rely instead on measures of such things as the structure and elaboration of student essays, the

sophistication of vocabulary, or the number of errors in grammar, usage, mechanics or style" (sect. 1).

Deane et al. (2013)'s philosophy that AEE represents one branch of a "division of labor" of scoring, rather than a scoring system that alone captures all or most of the relevant features of the writing construct, might be considered an emerging view of the AEE community. Indeed, as seen in the first chapter of this dissertation, Ellis Page considered AEE a replacement for human scoring, not a complement to it. A closer examination of Deane et al. (2013)'s methodology and results will serve to guide the hypotheses formed in the present study.

Deane et al. (2013) applied the cross-correlation methodology of using multiple rating systems over the same set of essays, to study *e-rater's* sensitivity to genre-specific writing skills. The purpose of the study was to measure how automated scoring would perform in a "writing from sources" assignment typical of the Race to the Top Assessments, a task which seeks to measure not only general writing qualities but strength of argumentation and accuracy of explanation—aspects of what might be called an argumentative genre. In 2009 and 2011, a convenience sample of 2,606 and 3,667 students, respectively, took either one or two CBAL (Cognitively-Based Assessments of, for, and as Learning) test forms each. Raters scored essays along a common rubric as well as rubrics specific to the test genres, and *e-rater* models were built to predict scores on the common rubric.

Then, to assess validity, Deane et al. compared (among other measures) correlations between the *e-rater* models and human scores along the genre-specific rubrics. In both datasets, the researchers found evidence to support the hypothesis that

different traits were being measured by human raters scoring along the common and genre-specific rubrics, which was then reflected in the correlations between *e-rater*, trained on the common rubric, and the genre-specific scores. In the 2009 administration, human rater correlations between common and genre-specific rubrics ranged from .75 to .86, and in the 2011 administration, human rater correlations ranged from .62 to .79. *E-rater* correlations ranged from .68 to .80 in 2009 and .61 to .78 in 2011. Deane et al. conclude that "the .64 cross-rubric correlation [between human raters along the genre and common rubrics] that we see in the 2011 Mango Street administration is small enough to support the conclusion that the human raters were responding to different traits" (Validity Evidence section, para. 5).

Deane et al.'s study incorporated the kind of predictive validity argument that the present research seeks to explore further. According to Deane et al., if *e-rater* captures the writing construct represented by the common writing rubric, it can be expected to stand in similar relationships to related measures of writing quality. For these authors, the common and genre-specific rubrics measure "different but related aspects of the same performance" and thus *e-rater* ought to be able to replicate the "moderate to strong correlations" that would be expected between the human ratings on each rubric (Data Analysis section, Validity Evidence subsection).

The present study can be considered a validation study in the vein of Deane et al. (2013)'s study, because it seeks to measure the strength of association of *e-rater*'s evaluation of common writing qualities with a rubric measuring higher-order skills. The present study, like others before it, incorporates multiple rubrics, as modeled in the Winters (1980) and Follman and Anderson (1967) studies described above. However,

instead of positioning the study as one that would evaluate the comparative validities of several different rubrics, it follows Deane et al. in assuming that what is in question is not the validity of the rubrics; these are implied to have been validated in some way or another (although validation of the rubrics is not explicitly discussed in that paper). Rather, like Deane et al., my study explores the degree of relationship between the construct captured by *e-rater* and other conceptions of the writing construct more commonly demanded from the college writing environment. Deane et al. wrote that "such predictive relations are not the only source of evidence, but are nonetheless critical in building a validity argument" (Data Analysis section, Validity Evidence subsection). Unlike these ETS authors, however, the objective of my study is not to build a validity argument for automated scoring in general, nor for *e-rater* in particular. It is, rather, to independently study the validity of *e-rater*, as both a summative and a formative agent for college writing.

Computer raters, in this conception, are hybrids: they function as raters, but also as internalized *rating systems*. If *e-rater* were treated as a human rater, the question would not arise as to whether a human rater could rate along two rubrics—researchers would simply train the person on one, and then on the next, rubric (as in the method described in Winters, 1980) and to the extent that a satisfactory level of interrater agreement were obtained, the rater would be considered trained on that particular rubric. But because *e-rater* is only able to directly measure a subset of writing qualities, and by virtue of its software limitations cannot simply be trained on all rubrics deemed appropriate for an assessment context, the proposal to compare its scoring to human scores on other rubrics makes sense.

Critiques of Timed Diagnostics

Since the study described in the next chapter cross-correlates scores on an essay administered under timed conditions, the final matter to be reviewed is the legitimacy of studying timed writing tests at all. Perelman (2012b) has argued that timed impromptu tests meet the definition of "bullshit" according to philosopher Harry Frankfurt's (2004) conceptual analysis. For Frankfurt, "bullshit" referred to discourse that is not concerned with the truth or falsity of a matter, but with effecting a particular outcome. By design, timed impromptus ask students to write on topics about which they have little knowledge (Perelman, 2012b, p. 430) and do not have the ability to research in the moment. Frankfurt, quoted in Perelman (2012b) notes that "bullshit is unavoidable whenever circumstances require someone to talk without knowing what he is talking about" (p. 430). Timed impromptus contrast with most college assignments, which revolve around the premise that students learn best when provided with rich material to read, discuss, and respond to over a matter of days or weeks. Impromptu writing without knowledge of a topic, contends Perelman, is unlike college writing in this fundamental way.

Relatedly, in a review of the Writing Assessment Program at Washington State University, Condon and Kelly-Riley (2004) uncovered an inverse relationship between assessments of critical thinking in timed essays and assessments of writing ability in these essays. These authors attribute that inverse correlation to the structural inadequacy of the timed exam, suggesting that the limitation of time, as well as potentially the "fact that students are long trained by various educational assessments in their K-12 schooling to consider timed writing in various reductive ways," serve to undercut the practice of inquiry (p. 67).

But if we accept Perelman (2012b)'s and Condon and Kelly-Riley (2004)'s objections to the adequacy of a timed impromptu to capture college writing and thinking skills, it is reasonable to wonder whether timed impromptus should actually be scored along college-based rubrics at all. Is the entire idea of evaluating timed impromptus for rhetorical competence, critical thinking, or argumentative skill quite meaningless?

Quite possibly, yes. A literature search of the correlation between the SAT essay portion and a student's college GPA only yields a claim from the College Board that essay scores have a "positive" relationship with first year GPA and first year English GPA (Shaw & Kobrin, 2012, p. 9). Providing an imprecise term such as "positive" rather than a specific correlation coefficient could imply that the relationship is in fact rather weak. Certainly, the process movement in pedagogy holds that the ability to research a subject and have time to write, read, and revise one's writing would be expected to produce a more nuanced writing performance, and probably a performance with a greater likelihood of creating knowledge, than a timed essay. So from the perspective of producing a response that yields knowledge and rhetorical effectiveness, the critical essay is superior—justifying its honored place in the composition classroom.

However, while Perelman's conceptual case and Condon and Kelly-Riley's empirical findings have merit, neither case indicates that there are *no* statistically significant differences in critical thinking, or in rhetorical effectiveness, among writers of timed essays. Maybe timed essays are indicative of some skill (however questionably related that skill may be to critical thinking). For instance, although students writing a timed impromptu will not be able to cite sourced data (aside from whatever the writer knows from memory) they will be able to produce narrative arguments drawing on their

own personal experiences to support the positions they take. Writing subjects will have to create reasons to support their positions—they will have to engage in acts of invention to examine subjects, construct or find lines of arguments, and identify materials to develop texts (c.f. Lauer, 2004). An author may present an appeal to intuition, or an argument from first principles. My study imagines that some students may be more able to produce more highly-skilled results from their process of invention than others. Therefore, the rubric created for this scoring session and the instructions given to the readers acknowledged that evidentiary support would not come from cited sources but could take the form of personal narrative (see next chapter). Thus, a compelling narrative may count as supporting evidence, and as evidence of critical thinking.

But on the other hand, if the scoring process yields categorically low scores indicating lack of proficiency, this would point to evidence that critical thinking and rhetorical skills do not manifest themselves readily in timed essays. As long as the study can find anchor essays that represent each score on the scoring scale, the distribution is free to be skewed toward the right, meaning that the concentration of scores are on the lower end.

To summarize: the present study is sympathetic to Perelman and other critics of timed essays as poor yardsticks for college writing success. At the end of the day, timed impromptus can only build the plausibility for any particular position taken by the author; they cannot ground that plausibility in an established discourse or in settled fact because of the restriction on sourcing. They are at odds with much of what the process movement is about. The kind of inventive process a student undergoes in a timed setting does not replicate that which is undergone during the course of a semester. But they are here,

especially in the form of placement essays for incoming first-year college students. They are a part of the essay libraries of automated systems hoping to proliferate. For these reasons alone, it is important to examine their validity.

Conclusion

For Ellis Page, what it meant to get the scoring "right" meant to agree with human scorers as often as they agreed with each other. It was irrelevant (indeed, expected) that computers would use different mechanisms, different proxy variables, to generate scores, than human readers. Since then, companies such as ETS have taken strides to improve both human agreement percentages and machine-human agreement. The move toward greater reliability has coincided with what the composition community has decried as highly invalid, artificial testing circumstances. (I have not tried to propose causal links between the increasingly asphyxiating standardization and greater reliability; however the conjunction between the increased reliability and increased standardization is notable.)

While the discourse around AEE remains polarized around what it means to get the scoring "right"—whether validity should be sacrificed upon reliability's altar, so to speak—two questions come to mind for composition studies. The first is whether AEE can capture, indirectly through exploiting cognitive capacity, college writing ability. The second is whether the reigning analytic paradigm of the Pearson correlation and related indicators of agreement is the best or only way to characterize what happens when computers score essays.

The study I undertook—described in the next chapter—compares *e-rater*'s performance in scoring a sample of 108 student essays (written to a *Criterion* prompt) to human scoring using a rubric inspired by the Framework Outcomes Statement. Data

analysis shifts adds a further analytic frame to the typical agreement metrics: an error analysis mode that begins to pull apart the monolithic character of presenting only aggregate data of agreement.

4. COMPARING RUBRICS: A CROSS-CORRELATION STUDY OF *CRITERION*®

Introduction

As the preceding literature review saw, the ascension of reliability (and correlation as its marker) to measure the legitimacy of AEE followed, logically and genealogically if not etiologically, from the symbolic engineering of AEE as systems whose communicative outputs should match that of humans.[9]

Woven into this litmus test was the assumption that mechanisms for getting to these outputs were irrelevant, as long as the outputs matched. As we saw, critics charged that not only does that matching occur only in contrived and specific circumstances, but that these circumstances fail to create ways of measuring writing skills validly—of representing a robust writing construct.

These positions are not mutually exclusive, and in fact often travel together as the operationalizations demanded by reliability may decrease the contexts of its validity. Timed impromptus enjoy a number of these characteristics: highly general prompts thought to increase fairness by reducing the importance of specific content knowledge (such as the maligned "is failure necessary for success?" prompt); timed conditions leading to relatively larger variance in response length that can serve legitimately or not as generative of ratings assignment; scoring rubrics restricted to formal characteristics rather than messier but important concepts such as "voice" or "quality of argument"; and raters financially incentivized to reach score consensus. Such reliability-enhancing

[9] (I say, "logically and genealogically if not etiologically" to acknowledge the prevalence of statistical correlation as a measure of legitimacy in many other domains where no "gold standard" exists. I do not mean to deny other historical inputs for this method's ascension.)

features create challenges for validity claims.

The present study replicates a typical context in which computer scoring might be employed: to rapidly assess first-year students. Two adaptations to the typical scoring paradigm have been made. First, in the scoring dimension, humans scored the essays on a Framework-derived rubric rather than a common rubric. Understanding how closely *Criterion*'s scoring captures higher order features of writing, such as those demanded of students in typical college settings, is an important step in assessing the mechanism by which automated systems may (or may not) improve student writing or be used to place students in appropriate writing courses. Second, on the analysis dimension, a frame of student impact was used alongside typical agreement measures to evaluate the software. If institutions want automated placement or evaluation to serve all students, a proposition to which most administrators will assent, attention must be given to human-machine score disparities, not only statistical agreements and correlations.

Note that this study examines the *Criterion* system's ability to model scores derived from human raters who are evaluating papers based on a rubric that measures Framework Outcomes Statement skills, as these represent skills that college teachers find important, at least according to an influential representative body, the Council of Writing Program Administrators (CWPA). While this research will not directly answer the question of whether students' *Criterion* feedback would iteratively improve writing, it would help shed light on the strength of the connection between the traits *Criterion* evaluates and those valued in the composition classroom. Some instructors report that simply the act of receiving instant feedback motivates students to submit more drafts, thereby improving outcomes regardless of construct validity (A. Klobucar, personal

communication, June 20, 2013; K. Morin, personal communication, October 11, 2013).

Research Hypotheses

Despite the majority of AEE trials having been done with standardized testing essays, some automated software applications have been marketed for K-12 and college classrooms. Take the Educational Testing Service's *Criterion* system, for instance. *Criterion* is an interactive technology that scores essays using ETS's *e-rater* scoring system, provides feedback to students, and enables web correspondence between teachers and students. Students can upload an essay to the system, click "submit" and be directed to a screen that gives an instantaneous score of 1-6, together with both default bullet points and the ability to click to different screens with more detailed analysis of organization & development, mechanics, grammar, usage, and style. As an illustration, the default feedback for a score of 4 is provided below in Figure 4.

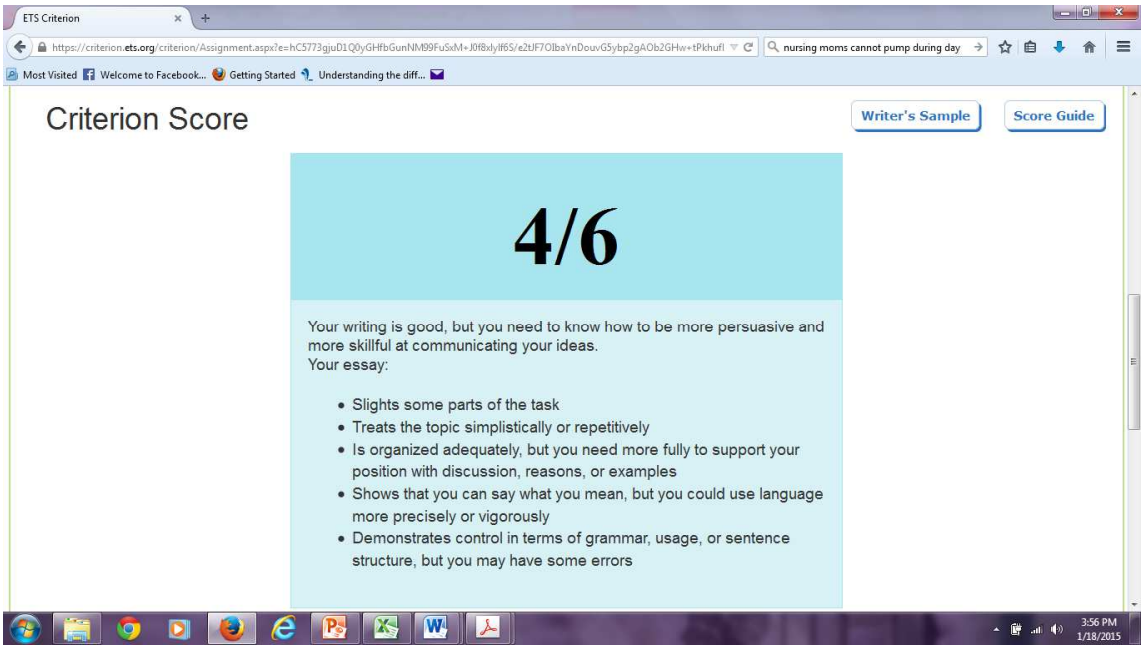


Figure 4. Screen capture of Criterion feedback page.

Bob Broad (2006) described opening his mailbox in 2003 to find a postcard from ETS, which he inferred had been sent to everyone on the National Council of Teachers of English mailing list, inviting him to learn more about the *Criterion* Online Writing Evaluation service. "How long does it take you to evaluate an essay?" the postcard asked. "Instantly…," it replied, if one uses *Criterion* (p. 223). In another study, Lim and Kahng (2012) reported that *Criterion* rated 5.8 million scripts in 2010.

The present research study will fill gaps in the knowledge about the role that writing evaluation software can play in college composition programs, where "higher order" writing skills such as critical thinking and rhetorical understanding often take precedence over grammar, mechanics and other features that are more readily assessed through software. Since the extent to which current AEE models can capture the college writing construct or constructs in the naturalized environments of classroom teaching or program-wide assessment is poorly understood, this study takes up the validity of AEE in college classrooms.

As implied, this project focuses on one particular software application because of its relative popularity and availability: the *Criterion* Online Writing Evaluation service by the Educational Testing Service (ETS). *Criterion*'s online service, about which more is available here http://www.ets.org/criterion, is a subscription-based online instructional tool somewhat like a simplified course management system that lets teachers create assignments that students can view and complete, that lets students see instructor feedback, and that contains a writers handbook. Differently from a course management system, though, *Criterion* contains the *e-rater* scoring engine.

With *e-rater* comes a writing topic library with more than 400 essay prompts that

can be selected as writing assignments, for which the scoring engine will provide a score along with default contextualizing remarks. As a second option, instructors who want to use *Criterion*'s scoring functionality can create their own prompts to be scored by *e-rater*. Such "scored instructor topics" are supposed to be written according to guidelines provided by the web service. Like the default prompts, the scored instructor topics come with a 1,000-word limitation. *Criterion*'s NLP software analyzes the prompt for purposes of providing relevant feedback and has a limit of 1,000 words for generation of holistic scoring. As a third scoring option, the system can be set to the "text edit" mode where essays longer than 1,000 words can receive diagnostic feedback but no score. (The system will generate an "advisory" when an essay is longer than 1,000 words. According to the online site, essays longer than that stand a chance of receiving an "unreliably high automated score" (https://criterion.ets.org/content/Advisory%20Table%2008-17-13.htm). (The ETS team itself, by virtue of their responses to my inquiry, would seem to support the finding that AEE's reliability is higher for essays under 1,000 words. One correspondent from ETS asked about my "motivation to investigate machine scores for responses greater than 1000 words, given that the tool/machine is best de[s]igned to evaluate essay type responses for timed tests?" (C. Ramenini, personal communication). Recently, *Criterion* added a fourth assignment type related to Common Core standards, but the version tested here did not present that option.

Three research questions follow from the three available scoring modes:

1. How does the *Criterion e-rater* scoring compare to human-assigned scores using a college-appropriate rubric on prompts from its database?

2. How does *e-rater*'s scoring compare with human-assigned scores using a college-appropriate rubric, for "scored instructor topics"?
3. How does *e-rater*'s scoring compare with human-assigned scores using a college-appropriate rubric, for longer (>1000 words) essays?

Due to constraints of time and funding, only research question 1, the scoring of answers to the *Criterion* prompt, has been completed to date and will be reported upon here. The other two questions have *not* been investigated as of the present time.

To operationally gauge the comparison, both measures of agreement and measures of distributional difference will be used. In the first category are: absolute agreement, adjacent agreement, Pearson correlation, Cronbach's alpha, and quadratic weighted kappa. In the second category are mean, standard deviation, and variance. All are standard measures in the field of educational assessment (Shermis & Hamner, 2012). The tests are further described in the "Data Analysis" section of this chapter.

What hypothesis might be generated for a cross-correlation study of research question 1? Given the literature discussed in chapter 3, including Congdon and Kelly-Riley (2004)'s finding that critical thinking was inversely related to writing ability in the tests administered by Washington State University, we might expect that the human readers, following the same scoring rubric, would score more similarly than they would to *e-rater*, with its immutable, pre-calibrated weights of formal features or "common" features of writing. In other words, we would expect more consistency from the scorers of the same rubric than we would between different rubrics. Here is one way to formulate this hypothesis:

H1: When humans score essays on a college-centered rubric—the Framework characteristics described herein—as opposed to standard AEE criteria, a significantly lower human-machine reliability will be observed than typically reported values (~.85 correlation as reported in Deane et al, 2013). (Scores will show low or, at best, moderate relationships per the statistical tests mentioned above.)

A finding of H1 would imply that computer scoring is less appropriate to the college classroom than hoped for, at least under circumstances where instructors and assessors wish to measure writing qualities such as rhetorical effectiveness or critical thinking.

If it turns out that the agreement measures and the distributional data and other measures of inter-rater reliability are not significantly different between the humans and *e-rater*, one implication would be that computer scoring *could* have limited application in the college classroom. Even though *Criterion* does not directly measure higher-order traits, an argument might be made that the scores themselves could be useful for students wishing to get a sense of how their essay might fare when graded by the instructor (or for teachers wishing to give students ballpark ideas of their work). Administrators could use *Criterion* for placement. It would suggest that whatever construct AEE captures is closely related to other writing constructs. Here is the null hypothesis:

H0: Machine-human reliability using *e-rater* will not differ significantly from human reliability using the Framework rubric.

One challenge of the present research design, though, is that a result in favor of H0 could be explained by several rationales, some supporting and some undercutting the argument that AEE scoring is valuable in college. The theory of cognitive capacity presented in chapter 3 predicts that stronger writers will have stronger across-the-board skills, while weaker writers will display weaker skills across the board, so that even rubrics that measure divergent aspects of the writing construct would still correlate highly with each other. Under this rationale, we might expect there to be little difference between *e-rater* scores on the common rubric and human scores on Framework characteristics for the same essay, since writers who have more control over surface features will be able to devote more effort to solving higher-order demands, whereas writers with less control will experience the reverse.

Thus, if H0 is found, the theory of cognitive capacity may be implicated. If H0 were to be borne out, however, it would be useful to have some initial evidence to discriminate among the various other rationales that might explain it. Understanding whether instructors can reliably score higher-order skills, whether timed impromptus facilitate the use of such skills, and whether these skills are coupled with lower order skills would contribute to a number of prominent debates in the field. Here are some other rationales that might explain a finding of H0:

(a) **Readers score based on surface quality regardless of rubric**. Perhaps college instructors cannot identify rhetorical knowledge or critical thinking as naturally as we might think they would be able to; thus, teachers and the computer software will

agree on writing quality because what is being evaluated in both cases are primarily surface features of writing.

(b) **Surface features are linked to overall quality.** Even if instructors may indeed be trained to reliably identify rhetorical knowledge and critical thinking in student essays, perhaps the constraints of the timed impromptu dictate that surface features such as essay length will nevertheless legitimately explain the lion's share of the score variance.

Based on the findings in the literature about the importance of this variable (see chapter 3), this study will track overall essay length in number of words as an independent variable. If divergences are seen between human and machine score correlations with overall word count, it may be tentatively concluded that human raters are able to review even timed impromptu essays for a richer feature set than computers do currently. This finding will be elaborated on in the "Results" section.

Methods: Data Collection

Beginning in January 2014, 168 students were recruited from 13 sections of two different first-year writing courses. The students consented to forward to the researchers several sets of essays written during the semester running from January-May 2014 (the precise number of essays each student forwarded depended upon the instructor's assignment sequence; essay that did not have a primarily argumentative focus were not used, for example.) The selected sections were located in two universities in New

England and comprised a convenience sample based on my own acquaintance with instructors and writing coordinators in those two institutions.

The two institutions serving as recruitment sites were Southern Connecticut State University (SCSU) (n=118; nine sections) and the University of Rhode Island (URI) (n=50; four sections). The first institution, SCSU, is a comprehensive university whose 11,000 students come predominantly from Connecticut, and which is part of the Connecticut State Colleges and Universities consortium. The second institution, URI, is a land-grant public research university of 16,000 students. Both institutions have first-year writing programs that reach the overwhelming majority of the student population. The participating students were enrolled in either a course at SCSU called ENG 112—Writing Arguments or a course at URI called WRT 104—Writing to Inform and Explain. The SCSU course is required of all students attending; the URI course is not required itself but is a popular option to fulfill required general education studies in writing. I was able to recruit five instructors at SCSU and two instructors at URI, each of whom permitted me to recruit students from their sections. (The timing of IRB approvals from both institutions facilitated my recruiting more classrooms at SCSU than at URI.)

The methodological value of incorporating more than one institution into the study is, first, that a multi-institution method can be expected to enhance reliability of the essay scoring, since no sections were scored exclusively by raters from their originating institutions, minimizing institution-dependent effects of raters (for further discussion of multi-institution studies in composition, see the website for The Citation Project at cite.citationproject.net). Second, the multi-institution design stood to produce a richer set of responses, given the distinctive student populations at the two schools.

In terms of the writing prompt assigned: *Criterion* has a library of 400-plus essay topics that teachers can assign students, on which *e-rater* has been trained to generate scores. In other words, *e-rater* comes with pre-weighted feature sets for the prompt based on previous training (as opposed to training up specifically on sets of essays provided by end users). Consonant with what we know about where automated scoring works best, ETS literature provides suggested response times and states that scoring is generally better in the timed environment. The topics are staged at different "levels" from fourth grade up to a level called "college level II." The prompt selected for this study is given in Figure 5:

Liberal Arts Education (Persuasive)

In the ancient world, the term "liberal arts" referred to the education appropriate for free people (as opposed to slaves). In modern American higher education, the term is used to describe an education that focuses on general, rather than vocational, knowledge. Proponents believe that a liberal arts education is valuable because it prepares students for life by teaching them how to think. Opponents contend that the study of topics unrelated to one's professional path is a waste of time. Is a liberal arts education worthwhile? Develop your position by using evidence from your own experiences, observations or reading.

Figure 5. *Criterion* prompt used in study.

This prompt appeared in *Criterion*'s college—level I (first year) library of prompts. Besides its surface appropriateness for the population sampled in this study, it was chosen out of my anecdotal knowledge that the value of education is a common

discursive theme in first-year composition; thus, an assignment about liberal arts education might be more amenable for teachers to fold into their curricula than other prompts. In the end, this intuition proved somewhat true, as two of the seven participating instructors stated to researchers spontaneously that this impromptu fit in thematically with their course content.

Per the ethics of researching student populations, students' participation was optional and instructors were not aware of which students took part, nor could they incentivize participation. Also, because of the practical need to work within participating instructors' existing curricula and syllabi, instructors were offered the option of assigning the *Criterion* liberal arts prompt as (1) in-class writing assignment via paper and pen, (2) an in-class assignment via computer, or (3) a homework assignment via computer. Those assigning the work as homework asked students to allot only 45 minutes for completing the assignment. Since the assignment was not tied in with course credit or course grades, there is no reason to believe students would have spent more than the recommended time on the essays, but the time-to-completion was not officially tracked.

With the exception of two of the 13 sections, all students completed the impromptu assignment within 30 days of the start of the spring semester (by 2/10/14 for the semester beginning on 1/13/14 for students at Southern Connecticut State University, and by 1/29/14 for the semester beginning on 1/22/14 for the students at the University of Rhode Island). Students in the two outlying sections, at URI, completed their essays within 60 days of the start of the semester. All participating students signed consent forms. Participants were assigned unique identification numbers and all personally identifying information was removed from essays.

Of the total 168 students signing informed consent documents at the outset, 106 students ultimately provided responses to the *Criterion* prompt. The 106 *Criterion* essays are the only set that has been scored at the time of this writing. All 106 essays were included in the sample plus two additional essays described below for a total of 108. While there exists the possibility of a selection effect in terms of which students returned essays and which did not, this effect would not compromise the study, whose goal is simply to compare machine and instructor scoring on a set of essays, not to draw conclusions from that sample to the larger populations at URI or SCSU. Therefore, self-selection was not a concern for this particular project, as it might be for another assessment about student learning that involved sample collection.

Because of the probability that essay word count (i.e., length) is related to the quality of a timed essay, this study experimentally manipulated the length of several professionally-written essays in order to test whether raters would be able to discriminate among essays that were short but well-developed (SD), long and well-developed (LD), short and under-developed (SU), and long but under-developed (LU). (The notions of "short" and "long" were left untheorized except to mean shorter and longer than the sample mean length.) If raters can conclude that shorter essays can still have developed themes and sophisticated rhetorical moves, and that longer essays can lack these features, we have evidence that essay length and quality do not have to be inextricably linked. The rationale for using a professionally-written piece by a noted scholar was simply that its relative competence as a piece of academic writing when placed alongside the student-written, timed essays had face validity.

Originally, three essays were to be manipulated: a 2,127 word essay written by Nannerl O. Keohane for the *Chronicle of Higher Education*, another essay by Bart Giamatti while he was the president of Yale University and a final piece by Robert Geer, the dean of SUNY-Albany's college of engineering. The challenge for the manipulation, however, was the difficulty in operationalizing the concepts of "developed" and "under-developed." The modifications needed to be inconspicuous enough so as not to tip off readers. Would repeating sentences be an acceptable method of creating versions that are long but under-developed? If that tactic were found to be too obvious, what about modifying words in the repeated versions? If so, how much content should then be modified? And if modifications are extensive, would at some point the modification not help to rhetorically "develop" the essay? In future trials, work needs to be done to tie the manipulations to better-defined rubric qualities.

Ultimately, the Geer essay was the only one of the three manipulated essays to be used as part of the study. The Keohane and Giamatti essays were so long relative to the students' papers that cutting them sufficiently in order to create the SD and SU versions became a challenge. The Geer essay, however, was short enough that an even-shorter version was still comprehensible. I created a version of this essay that was significantly shorter than the mean word count of the student essays (m = 353 words) to represent the SD version and one that was longer, but achieved its length through repetitive phrasing rather than the addition of evidence or novelty to represent the LU version. The word counts of the manipulated essays were 274 and 382 words, respectively. If the teachers were able to take into account higher-order values during their scoring, they should have

scored the shorter but developed essay higher or as high as the longer but not more developed essay.

Methods: Data Preparation

Essays from the five sections that were hand-written were transcribed into typewritten form. Efforts were made to preserve the original structural and mechanical elements of the essays, including placement of paragraphs, spelling, punctuation, capitalization, and so forth. The reasoning for the decision to have instructors score typewritten transcriptions rather than original versions was to preserve parity between the form graded by machine and the form graded by instructors—both would receive the printed version of each essay. Again, in an assessment that compared students to one another, or classrooms to one another, the existence of multiple test forms would threaten validity because the typed essays would have enjoyed the advantage of spelling and grammar checks not to mention easier editing than the paper form administrations. Although evidence conflicts as to whether typed or handwritten examinations enjoy any sort of scoring advantage (Mogey, Paterson, Burk & Purcell, 2010), transcribing the paper administrations eliminated this possibility. (Note that transcriptions preserved original spellings and punctuations.) However, the purpose of this study was not a comparison of subsets of essays; it was a comparison of scoring procedures, so the variation in administrations was not germane. The only distinction between the essays that the people read and those scored by *e-rater* was that some students added titles to their essays, which the human readers would have seen, and *Criterion* would not have. The user interface instructs users to leave off titular content because titles may, at the risk

of anthropomorphizing, confuse the scoring algorithm.

In preparation for scoring, the essays were assigned to the five participating raters using a random number generator. Randomization was done to ensure that no rater would receive an unequal proportion of essays from any one classroom and there was no systematic pairing of raters (see Winters, 1980). No rater read essays from her own classroom, and this was ensured during the assignment phase.

To ensure consistency of scoring, a training set of nine student essays was assembled from the student sample; the training set comprised six canonical 1-6 score essays (not in that order) and three samples that represented borderline 3/4 splits. (Scores were based on the scoring model developed for this exercise, described later in this section.) The training recommended by Perelman (personal communication, May 13, 2014) was to use 12 essays taken from the sample set (consisting of canonical, or definitive, 1-6 scores in the first packet, followed by a second packet with a high-scoring shorter essay, low-scoring ESL essay, and borderline 3/4 split, then a second borderline 3/4 split, low-scoring longer essay, and high-scoring ESL essay in the third packet). This training only utilized nine essays because after scanning approximately 50% of the sample, it was determined that the sample size was too small to yield canonical examples of ESL papers, high-scoring short, or low-scoring long papers. Canonical papers must have as clearly discernable relevant attributes as possible and this was not found. Even finding canonical "1" and "6" scoring essays was also a challenge given simple probabilities. If the 106 essays were normally distributed, with a mean, say, of 3.5 and SD of 1, then 95 percent of the data would fall between 1.5 and 5.5, meaning that only five or six essays would be 1s or 6s. On the other hand, we could expect a good

proportion of 3s and 4s, making a closer look at the "upper half-lower half" border line important.

In a second minor deviation from standard scoring practice, reader scores on the training set were included in the overall dataset of scores. (Readers had already been assigned to all essays before the training set was developed and scores on the training set from the assigned raters were taken for the dataset.) Typically the training set is of course not included in the overall dataset that is then analyzed statistically, since reliability should only be calculated once the raters are trained. However, in this case, working with as large a dataset as possible was appropriate for a pilot study where the scores had no impact on students or institutions and would outweigh any concerns that including training essays could lower inter-rater reliability.

Developing the scoring guide. A thorough literature review did not reveal any published studies on rubrics that had been derived from Framework outcomes, so it was not possible to examine any existing reliability or validation data concerning this kind of rubric. Yet, as Schwegler (2014) has noted, the Framework outcomes provide "a set of traits that instructors look for in compositions" (R. Schwegler, personal communication, May 23, 2014) much in the same way that the Diederich scale of features provided this set of traits in Diederich's factor analysis of what secondary school teachers look for.

Because of the shortage of reported studies on using the Framework for assessment, this project began developing the scoring rubric to be used in the study by building upon a pre-existing portfolio-scoring rubric developed by the University of Rhode Island (URI)'s Department of Writing & Rhetoric for scoring 100-level end-of-semester portfolios, and based on the Framework. That rubric is shown in Table 2, below.

Reliability figures for that rubric were not readily available and it is also a rubric using only three score values (1-3), so there would be no immediately obvious basis for drawing conclusions about the 6-point rubric developed for this study, at least from simple agreement numbers.

Swain and Le Mahieu (2012) describe the process of using one rubric as the starting point for developing another rubric. A modified version of this process was followed in building the rubric used in this project. I consulted with Robert Schwegler, a faculty member involved in developing the original URI portfolio rubric, and with other experts including Chris Anson and Les Perelman. The guiding principle adopted from the URI rubric was that under the attribute of "rhetorical knowledge," the purpose should be clear, the voice ought to contribute to fulfilling the author's purpose, and the implied audience should be appropriate. In the case of the timed impromptu essays collected for this study, these criteria were interpreted to mean that the essay argued a position, that it was written in prose typical of academic writing (this feature was not further defined, but seemed to be something that reviewers understood intuitively), and that the voice or personality of the essay seemed to be working in concert with the purpose. (An essay written with a noticeably colloquial voice, for instance, may work against the purpose of arguing a position for an academic audience.)

For the attribute of critical thinking from the *Framework* Outcomes Statement shown in chapter 3, which was not explicitly employed in the URI rubric, the rubric here developed indicated that a strong essay would have provided strong and even insightful supporting data/evidence for the writer's position. Condon and Kelly-Riley (2004) described a six point, seven dimension analytic scale used to measure critical thinking at

Washington State University. The scale was claimed to yield reliability coefficients of .8 and higher (p. 60). The scale included measures such as identifying the problem at issue, the student's and other's perspectives, identifying key assumptions, assessing supporting evidence, considering the influence of context on the issue, and assessing implications (pp. 71-72). In order to incorporate some of these ideas into a holistic rubric, I chose to focus on presenting the author's position and providing strong supporting evidence. Expecting impromptu-writing students to explore the problem, context, and key assumptions seemed to be beyond the expectations of even strong papers.

In terms of conventions and craft, the rubric created for this study followed the URI rubric in using the litmus test of whether errors impede meaning. Doing so eliminated the mindset of "counting" mistakes that can otherwise infect scoring initiatives. The qualities of "reflective introduction" and "understanding of genres" were removed from the new rubric as they were not relevant to the essay assignment.

	EXEMPLARY	**PROFICIENT**	**UNACCEPTABLE**
Reflective Introduction	The reflective introduction demonstrates the writer's insights and thoughtfulness about his or her learning in this course and/or makes connections across projects and/or to other courses. The introduction offers a variety of support for claims about what the writer has learned, including evidence from the working folder or previous drafts. The writer addresses in detail why entries were chosen and/or how they were revised for the portfolio. The writer uses terms from the course content or materials comfortably or expertly (i.e., genre, audience, revision). The introduction successfully prepares readers for the rest of	The reflective introduction demonstrates some thoughtfulness about the writer's learning in the course; it offers support for claims about what the writer has learned and may make at least one connection to other learning. The writer addresses why entries were chosen or how they were revised for the portfolio. The writer uses terms from the course content or materials competently (i.e., genre, audience, revision). The introduction competently prepares readers for the rest of the portfolio.	The reflective introduction offers only one or two examples of the writer's learning or makes broad, unsupported, or generic claims about the class. The writer addresses only briefly why entries were chosen or how they were revised for the portfolio. The writer does not use terms from the course content or materials, or uses the terms in unusual contexts. The introduction appears disconnected from the rest of the portfolio.

	the portfolio.		
Rhetorical Knowledge	The writer has made apt decisions for each entry in terms of audience, purpose, and voice. The target audience for each piece is appropriate or consistent; in every entry, the purpose is clearly stated and fulfilled, and the voice in each case contributes to fulfilling the purpose.	The writer has made apt decisions in terms of audience, purpose, and voice with only one or two lapses. The target audience for each piece is generally appropriate or consistent; the purpose is stated and fulfilled, and the voice contributes to fulfilling the purpose.	The writer has not made apt decisions in terms of audience, purpose, and voice, or there are several lapses. The target audience for more than one piece is inappropriate or inconsistent; the purpose is difficult to determine or sometimes unfulfilled; the voice may or may match up to the purpose.
Understanding of Genres	Each of the three project entries illustrates a command of the intended genre and/or fulfills the criteria for the original assignment. For each genre represented, the writer has explicitly and implicitly demonstrated the features of the genre and has used those features to good effect (e.g., in analyzing a text, the writer has used evidence from the text effectively to achieve the goals of the piece).	Of the three project entries, at least one is very successful in illustrating a command of the intended genre or the criteria for the assignment; the other two entries may be on target but less polished. The features of each genre are present but are not manipulated to the best effect (e.g., in analyzing a text, the writer has tried using evidence from the text to achieve the goals of the piece but with mixed success).	Only one of the three project entries illustrates a command of the intended genre, or the criteria for the original assignment are fulfilled in only one entry. The writer has misunderstood or misrepresented key features of two of the three assigned genres.
Conventions & Craft	The reader has no questions about sources or documentation; there are minimal errors in usage, grammar, punctuation, or mechanics—none that impede meaning.	The reader may have one or two questions about sources or documentation; there are a few errors in usage, grammar, punctuation, or mechanics that may impede meaning.	The reader has a number of questions about sources or documentation; there are a number of errors in usage, grammar, punctuation, or mechanics that impede meaning.

Table 2. Example of URI rubric for portfolios based on Framework.

Based on the URI rubric presented above, the following scoring guide was created for this study (Figure 6):

Scoring Rubric for College Writing Based in *Framework* Criteria

The rubric reflects the degree to which essays: are written toward an academic audience, convey and fulfill the purpose of the assignment, demonstrate critical thinking, and contain minimal surface errors.

Impromptu essay prompts do not present "settled" debates. Lacking sources, writers may rely more on personal narrative and/or anecdotal evidence to support their position, rather than on "hard" data. Essays may likewise contain reasoning from first principles or appeals to intuition. Paper-and-pen writers will not

have had the advantage of spell checkers. Still, if well-constructed, strong essays should be able to offer compelling reasons for their position that contribute to an overall impression of rhetorical competence and effectiveness. Strong essays should be written in an academic voice and should be relatively free of major surface errors.

PROFICIENT	6 Exemplary	Essay's target audience is consistently appropriate; the purpose is clearly stated and fulfilled, and the voice contributes to fulfilling the purpose. Essay presents author's position and provides strong, even insightful supporting data/evidence. There are minimal errors in usage, grammar, punctuation, or mechanics—none that impede meaning.
	5 Good	Essay's target audience is generally appropriate; the purpose is stated and fulfilled, and the voice contributes to fulfilling the purpose with minimal lapses. Essay presents author's position and provides strong supporting data/evidence. There are only minimal errors in usage, grammar, punctuation, or mechanics that impede meaning.
	4 Acceptable	Essay's target audience is often appropriate; the purpose is stated and fulfilled but not as clearly as in 5-ranked essays, and the voice contributes to fulfilling the purpose most of the time. Essay presents author's position and provides some degree of supporting data/evidence. There may be a few errors in usage, grammar, punctuation, or mechanics that impede meaning.
NOT PROFICIENT	3 Developing	Essay's target audience is not always clear or appropriate; the purpose is not clearly stated and fulfilled, and the voice may not contribute to fulfilling the purpose. Essay may not fully present author's position and may have limited supporting data/evidence. There are noticeable errors in usage, grammar, punctuation, or mechanics that impede meaning.
	2 Emerging	Essay's target audience is generally not appropriate; the purpose is not clearly stated and fulfilled, and the voice does not contribute to fulfilling the purpose. Author's position is unclear and essay does not offer strong supporting data/evidence. There are significant errors in usage, grammar, punctuation, or mechanics that impede meaning.
	1 Needs Improvement	Essay's target audience is not appropriate and/or discernible; there is a lack of coherence in the essay's purpose and/or voice. Essay presents no clear position and lacks supporting data/evidence. There are substantial errors in usage, grammar, punctuation, or mechanics that impede meaning.

Figure 6. Rubric for Essay Reading Session

Methods: Data Scoring

On May 23, 2014, the rating session was held, with five raters participating. The seven instructors from SCSU and URI who agreed to invite their classes to participate in

the study were invited to serve as readers in the scoring session at the end of the semester. Ultimately, five were able to participate. Each rater was one of the instructors whose students took part in the study, and both SCSU and URI were represented by raters.

Following Winters (1980), a random assignment of essays to readers ensured that there was no systematic pairing of readers to each other, nor of readers to essays from any particular course section. All essays were read once then redistributed for second readings. Essays were coded according to institution, course number, instructor, and student. No instructor was permitted to score essays from his or her own course sections, per the informed consent agreement (Appendix 1).

Each of five participating raters rated approximately 22 essays in each of two separate rounds, the first round taking place in the morning and representing first reads of all 108 essays (106 student essays plus the one manipulated essay in SD and LU forms); the second taking place in the afternoon and representing second reads of the essays. This meant that all papers were read and scored twice in total. Scores were recorded directly on each printed copy of the essay; two copies were printed so as to avoid second readers inadvertently noticing what rating the first readers had assigned a paper. After scoring came to a close, each rater was given a $100 Visa gift certificate consistent with URI policy on incentive payments.

Methods: Data Analysis

Recall that the set of essays that were collected and scored at the present time consisted of student essays written on a topic selected from *Criterion*'s database, on the value of liberal arts education. The second set of essays collected (but not scored yet)

were responses to instructor-developed topics that are generally fewer than 1,000 words.

For the condition tested in this study, *Criterion* holistic scores for each essay were cross-correlated with holistic scores produced by the college instructors trained in the scoring of a rubric based on the Framework for Success (see Figure 2 in chapter 3). Many institutions besides URI draw explicitly or implicitly on the Framework, so using that rubric for all of the essays in the study would not be unfounded. Originally, the research design had incorporated a dual-rubric method, scoring the essays along the 6+1 rubric described in chapter 3, as well as a Framework rubric. However, with limited available funding, only one scoring session, using the Framework-derived rubric, was possible.

As mentioned, the essays were scored by human readers in May 2014. Scores were recorded into Excel over the summer of 2014. Splits—ratings that differed by more than one score value between readers—(n=24) were resolved by me over a roughly 24-hour period in August 2014. The split adjudication process was simply to review the rubric and re-score the 24 essays along that rubric. A Resolved Score (RS) for each essay was then generated. In the case of agreement or adjacent agreement (scores within one value between readers (e.g., a 2 and a 3) between the two readers, the RS was the average score. Following common practice, in the case of splits the RS was equal to the score assigned it when I resolved the splits. The rationale for not simply averaging split scores is that two raters' agreement is thought to further confirm a score than would simply one rater. If two raters do not agree, that suggests little confidence in one or the other of the scores (or both) so the essay should be re-scored.

Essays were then cut and pasted into ETS's online *Criterion* system during the fall of 2014, and each received a score. Four essays were returned as unscorable because

they were over 1,000 words. Data analysis was done using SPSS. Of interest are the mean, median, and modal score for each reader, as well as the standard deviation (a measure of how dispersed the scores are) and the shape of the distribution of scores (kurtosis and skewness). Descriptive statistics for Reader 1, Reader 2, the Resolved Score, and *Criterion* scores are as follows:

Reader 1		*Reader 2*	
Mean	3.46	Mean	3.7
Standard Error	0.11	Standard Error	0.12
Median	4	Median	4
Mode	4	Mode	4
Standard Deviation	1.18	Standard Deviation	1.22
Sample Variance	1.39	Sample Variance	1.5
Kurtosis	-0.5	Kurtosis	-0.3
Skewness	0	Skewness	0.06
Range	5	Range	5
Minimum	1	Minimum	1
Maximum	6	Maximum	6
Sum	374	Sum	400
Count	108	Count	108
Confidence Level(95.0%)	0.22	Confidence Level(95.0%)	0.23

Resolved Score		*Criterion*	
Mean	3.44	Mean	3.62
Standard Error	0.11	Standard Error	0.09
Median	3.5	Median	4
Mode	3.5	Mode	3
Standard Deviation	1.13	Standard Deviation	0.91
Sample Variance	1.28	Sample Variance	0.82
Kurtosis	-0.4	Kurtosis	0.18
Skewness	0.04	Skewness	0.21
Range	5	Range	5
Minimum	1	Minimum	1
Maximum	6	Maximum	6
Sum	372	Sum	376
Count	108	Count	104

Confidence Level(95.0%)	0.22	Confidence Level(95.0%)	0.18

Table 3. Descriptive statistics by reader type.

Mean scores ranged from 3.44 (RS) to 3.7 (R2). The median score for R1, R2, and Crit was 4 and for RS it was 3.5 (recall that RS is the average of R1 and R2 in the case of exact or adjacent agreement; otherwise, it is the adjudicated score). Modal scores were 4 for R1 and R2, 3.5 for RS and 3 for Crit. Some of the other statistical properties will be covered in the section, "Distributional differences."

Agreement Measures

A *correlation* is a statistical technique that provides a sense of how closely two variables are associated (Tabachnick & Fidell, 2013, p. 56). It is considered a measure of consistency. The Pearson product-moment correlation coefficient (represented as *r*) is the most commonly-used measure of correlation (Tabachnick & Fidell, 2013). A Pearson correlation ranges from +1 to -1. A correlation of 0 means that the variables are not associated at all, while a correlation of +1 means that they are strongly associated in a positive direction (as one increases in value, so does the other or as one decreases, so does the other), and a correlation of -1 means that the two variables are strongly inversely related (as one increases, the other decreases).

A correlation matrix (Table 4) shows the *r* values for the six possible combinations of variables: R1 with R2, R1 with RS, R1 with Crit, R2 with Crit, R2 with RS, RS with Crit (listed as R1R2, R1RS, R2RS, R1Crit, R2Crit, and RSCrit, respectively). The R1RS and R2RS correlations were the highest (.795 and .739) because

the resolved score was generally the average between the two reader scores (except for splits). The third-highest correlation was that between *Criterion* and the Resolved Score (.637). We would expect that the association between *Criterion* and the Resolved Score would fall in the moderate range, and indeed it does, because *Criterion* has of course been trained on a common rubric, and the readers were trained on the rubric incorporating Framework Outcomes Statement qualities.

Correlations

		R1	R2	RS	Crit
R1	Pearson Correlation	1	.497**	.795**	.589**
	Sig. (2-tailed)		.000	.000	.000
	N	108	108	108	104
R2	Pearson Correlation	.497**	1	.739**	.577**
	Sig. (2-tailed)	.000		.000	.000
	N	108	108	108	104
RS	Pearson Correlation	.795**	.739**	1	.637**
	Sig. (2-tailed)	.000	.000		.000
	N	108	108	108	104
Crit	Pearson Correlation	.589**	.577**	.637**	1
	Sig. (2-tailed)	.000	.000	.000	
	N	104	104	104	104

**. Correlation is significant at the 0.01 level (2-tailed).

Table 4. Correlation Matrix.

Perhaps the most unexpected finding was only a moderate, rather than a strong, correlation between the two human readers (.497). Operational standards at ETS require correlations above .70 (Deane, et al., 2013) and indeed the Shermis and Hamner (2012) study of multiple data sets featured human correlations in this range. While the training session here did not officially compare ratings, personal impressions were that raters agreed with and internalized the rubric. Moreover, because scores are not associated with

student performance, the scores are instructive for the purposes of studying efforts to apply a *Framework* rubric to timed essay questions.

While correlation will give a sense of how two datasets move together, it is possible for two arrays to be perfectly correlated but not agree on any of the data points (say one rater is uniformly more severe than another), so a sense of overall agreement is also needed—a measure of consensus. Exact agreement for R1R2 was 33 essays, or 30.55 percent. Exact+Adjacent agreement for R1R2 was 84 essays, or 77.78 percent. Virtually all of the human ratings (97.22 percent or 105 of the 108 ratings) were within two points suggesting some consistency of scores. Again while these numbers are not ideal in a high-stakes situation, they are not unheard of in other contexts. Leydens and Olds (2012) report on a study where 22% of the sample was adjudicated because scores were not in exact or adjacent agreement (p. 251), which would place agreement at 78%. Yet agreement alone is an inadequate measure since it does not account for agreement by chance.

Cohen's kappa is used for assessing inter-rater reliability when the measure is categorical in nature and provides a mathematical formula for discounting for the probability of agreement merely by chance. Weighted kappa is less punitive to disagreements that are closer together on the scale than those that are farther apart, which is useful because the interval data from essay scoring is thought to have some "underlying trait" that increases along with the scale, making weighted kappa a helpful measure (Shermis & Hamner, 2012, pp. 23-24). Quadratic weighting—a standard measure for interrater reliability in educational assessment—uses weights proportional to the square of the distance apart. The quadratic weighting has the effect of being more

generous to closer-together responses that either unweighted or linear weights. Kappa ranges from 0 to 1, with higher numbers indicating greater agreement. The quadratic weighted kappa coefficient for R1R2 was .4867. The coefficient for R1Crit was .5621 and the coefficient for R2Crit was .548.

Cronbach's alpha for R1R2 was .664. Cronbach's alpha is another measure of the consistency (internal) of a set of tests measuring the same construct. It can be used to measure interrater reliability (Stemler, 2004). Cronbach's alpha was not used on the *Criterion* set, because this coefficient decreases in accuracy when faced with missing data, and there were four unscorable essays in *Criterion*.

While I will not attempt more than a speculation as to the lower human inter-rater reliability than generally accepted for high-stakes scoring, several rationales may be offered: (1) the operationalization of *Framework* features such as rhetorical knowledge and critical thinking provided by our rubric was unsatisfactory; features were not clearly defined enough and as a result, readers could be expected to differ on their interpretation. Or (2), readers simply needed more training on the rubric. This hypothesis could be tested by rescoring the essays with new readers who had been trained more extensively on the rubric.

However, there is the possibility that the problem is due to construct-irrelevant variance rather than a problem with rubric interpretation or definition, because the second reads were done simultaneously, and directly after readers broke for lunch. It is possible that there was a rater severity drift after the break, which might explain the higher sample mean for the second reads (3.70 compared to 3.46). Increased severity before—and leniency after—a meal was found in a study of parole board hearings (Danziger, Levav,

& Avnaim-Pesso, 2011). Supporting this potential explanation of leniency, six splits were adjudicated in favor of the R1 score; three were adjudicated in favor of the R2 score; 11 were adjudicated as the middle score between the R1 and R2 scores; and four were assigned lower scores than either R1 or R2 originally assigned. Interestingly, all three scores adjudicated in favor of R2 were situations where the R2 score was actually lower than the R1 score for that essay. In other words, *none of the adjudications favored the higher R2 assignment*, suggesting that the R2 scores may have been relatively lenient. Only two scores were resolved in favor of the higher score anyway. However, if all readers were consistently more lenient after lunch, Pearson *r* would not necessarily be affected, so it may be the case that some raters' leniency increased while others did not. More tests of rater consistency may be indicated here.

Regardless of these limitations, given the correlation of .64 between the Resolved Score and *Criterion*, the study lends limited support to the expectation that machine scoring on a common rubric will be moderately but not strongly correlated with scoring along a rubric more focused on higher-order features of writing such as the focus, quality of evidence, and writerly voice. (Recall from chapter 3 that Deane et al. (2013) found correlations of .61 to .80 between the *e-rater* model trained on a common writing rubric, and three writing prompts where humans used a genre-specific rubric to score the essays.) The results also do suggest that one cannot expect *Criterion*—which has been trained to predict scores on generic feature sets—to score essays as a college instructor might.

However, more training is needed for the human instructors so that they will be able to score essays more reliably along the *Framework* criteria. Until such time, these findings must be interpreted with caution. But because R1 and R2 did not correlate as

closely as RS and *Criterion*, hypothesis H1 could not be confirmed from this study. Recall that H1 was framed as a comparative: the scores within the humans' rubric were hypothesized to correlate more closely than the scores across rubrics.

Covariance with Essay Length

While H1 predicted different inter-rater reliability levels between humans and machines but found similarity, the opposite trend was found in looking beyond reliability: differences have been found where similarities might have been expected. Further analyses were done to explore these findings. The first of this "difference" concerns the correlation between essay length and score that has been observed elsewhere in automated scoring studies. That trend was evident in this study as well: Table 5 depicts the correlations between overall word count and essay score as a function of reader. *Criterion* had the highest correlation, almost double that of the lowest correlation, Reader 2. The correlation of .539 of human Resolved Score and essay length equals a shared variance (r^2) of 29 percent (.539 * .539 = .29). Shared variance can be interpreted as the percent of one feature attributable to another.

R1Length	0.458
R2Length	0.424
RSLength	0.539
CritLength	0.836

Table 5. Score-Length correlations by reader.

Reprinting the figure from Perelman (2012a) (Figure 7), the finding of a shared variance of 29% fits exactly with the trend that he uncovered: the length/score correlation

is stronger the less time a student has to write an essay. (Perelman's chart ranges from 20- minute standardized assessments to placement examinations at MIT allowing three days for completion). The essay I assigned had a 45-minute time limit, which places it between the blue (25 minute) bars and the red (one hour) bars. As the arrow denotes, the shared variance of 29 percent clocks in between of those observed measurements.

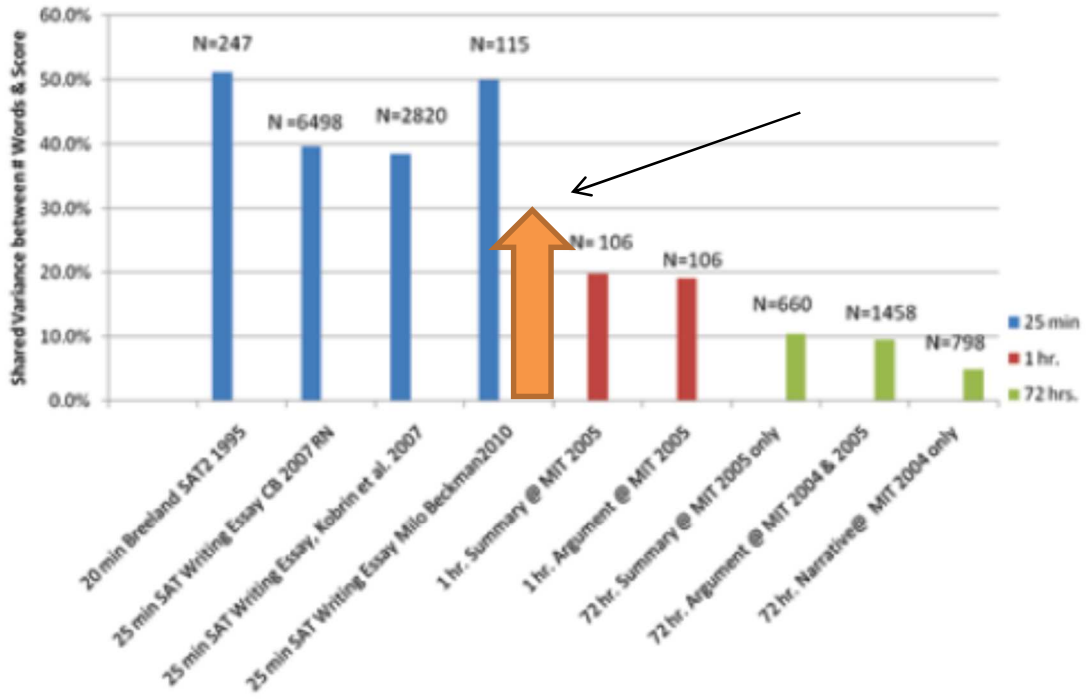


Figure 1. Shared Variance between Holistic Score and Length as a Function of Time Allowed

Figure 7. Reprinted from Perelman (2012a) with orange arrow depicting present study results.

That the RS shared variance appears to fall along the same trend line that Perelman has discovered is somewhat unexpected, in the sense that essay readers in this study were provided ample time to score essays, with the hope that they would not default to visual cues such as length in making their assessments, and were provided training essays in which the canonical "6"-scoring essay was shorter than the average

essay in terms of number of words. This training attempted to invite readers to jettison any preconceived notions they had about scoring these off-the-cuff assignments, without on the other hand biasing results by explicitly instructing readers to ignore length as a factor. It happened that at least one participating instructor had scored standardized tests for a testing company in the past, and so the message that form ought to be subordinate to content was a meaningful one since it did not hold true in her previous scoring experience.

Consequently, the resulting moderate correlation of length and score lends limited support to the idea that the length-score association is not solely an artifact of using a standardized rubric or of other limitations of scoring standardized examinations. In other words, the association of length and score is more likely attributable to the moderate likelihood that longer essays are indeed better-developed than shorter ones, especially when time constraints will produce a larger range of lengths. See below, Table 6.

Descriptive Statistics: Essay Length	
Mean	352.86111
Standard Error	14.676973
Median	315
Mode	271
Standard Deviation	152.52758
Sample Variance	23264.663
Range	935
Minimum	78
Maximum	1013
Count	108
Largest(1)	1013
Smallest(1)	78

Table 6. Descriptive statistics for variable of Overall Word Count.

As the table shows, the range of essay length was 935 words (from 78 to 1013 words). So, the longest essay in our study was nearly 13 times longer than the shortest essay. Essay assignments with longer time limits are unlikely to produce essays with this large a multiple between shortest and longest, because slower writers will be able to "catch up" with faster ones to reach what is either specified or understood to be a reasonably apportioned response (one would not expect see, for instance, a range of 3 to 39 pages for even a 72-hour placement exam). If we want to test the hypothesis that the length-score association is explained by the inverse relation between word count variance and time allotted (length varies less with longer writing duration), we would need to have multiple essay sets such as the ones in Perelman's study. We would then obtain Pearson *r* coefficients for length-score on each set, as well as measures of variance such as range and standard deviation. Then, we would plot the Pearson *r* coefficients against these measures of variance to see whether the essay sets where length varied the most were also the ones where length mattered the most. Unfortunately, this is outside the scope of the current study.

Moving from the RS length-score correlation to *e-rater*, the shared variance between score and essay length for the *Criterion* scores was much higher than the human reported variance, at 70 percent (.836 * .836). A higher shared variance does not necessarily mean that the computer was *explicitly* counting words, however. Indeed, Chodorow & Burstein (2004) noted that *e-rater* does not contain a direct measure of essay length in its feature set, an assertion corroborated by the Construct Decomposition chart provided in chapter 3 as Figure 3. But it does mean that essays that are longer have a strong tendency to be scored more highly by *Criterion*—this association is stronger, in

fact, than *any* of the observed relationships between length and human score reported by Perelman (2012a) and reprinted above. As the current research shows, essay length is not itself causative of high score, since the associations found were significantly lower.

While Perelman and others seem to suggest that any essay condition where length is the overwhelming factor in essay quality is illegitimate, one might respond that longer essays are, quite often, simply better—presumably because they constitute more developed responses. Indeed, as I have said some contribution of essay length to score is probably logical—a five-word essay, a 50-word essay and a 500-word essay are going to represent quite different rhetorical performances. *However, this study demonstrates that essay length need not be as strongly connected with score as automated scoring might indicate.* Again, *Criterion*'s feature weights were set during prior calibration sessions with human-scored essays, so the implication of the present project is that there are different ways to score even timed impromptu essays such that length is not the overwhelming feature taken into consideration. The limitation of this study is, of course, that once the parameters of scoring are loosened up, reliability decreases.

Distributional Differences

Beside the discrepancy of length-score correlation, the second notable distinction is the distribution of score frequencies between the human and machine scoring. As seen in Figure 8, the superimposed score distributions for *e-rater* look graphically quite distinct from those of the human readers. (Figure 9 shows the individual histograms.) Specifically, *Criterion* assigned the *fewest* number of 1s, 2s, and 6s, but the *most* number of 3s of all the scorers. The number of 4s and 5s assigned by *e-rater* fell within the range

observed in the human rater assignments. Exactly 80 *Criterion* scores were either 3s or 4s, as the figure indicates—more than three quarters of the sample.

Note that, as discussed earlier, RS is the average of R1 and R2 except in the case of splits, where RS takes on the value of the adjudicated score. So RS can be a half-score value. Half scores such as 1.5 or 2.5 were included in the bin for the lesser of the score values. Because of the way resolved scores were assigned, the unique histogram for RS does not appear normally distributed or at least appears skewed—there are more 1s than 1.5s, more 2s than 2.5s or 3s, and more 5.5s than 5s. One would expect more whole numbers than fractions, because the 24 adjudicated scores were assigned a whole number, but the explanation for the larger number of 2s than 3s remains to be seen. Possibly, the adjudications had the effect of lowering the scores. One explanation is that, as mentioned, the 24 adjudicated scores were lower than the R2 scores. Analysis of the adjudicated score distribution could add clarification.

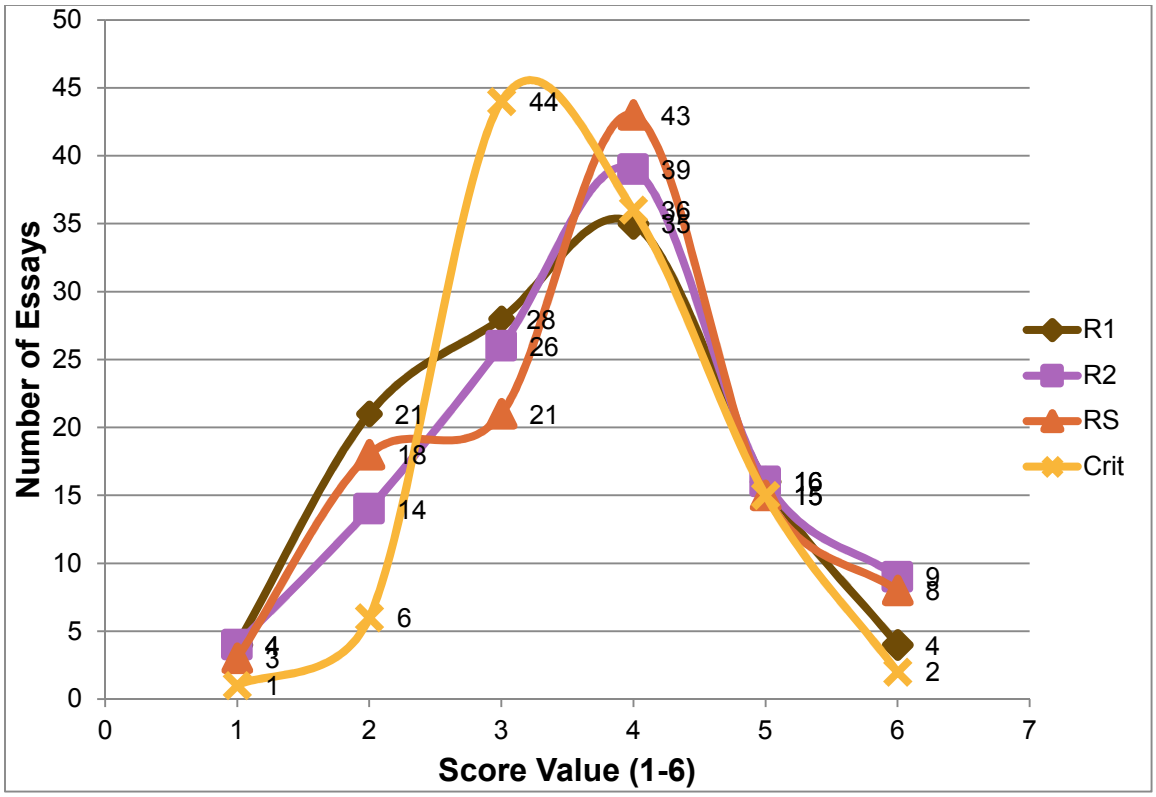
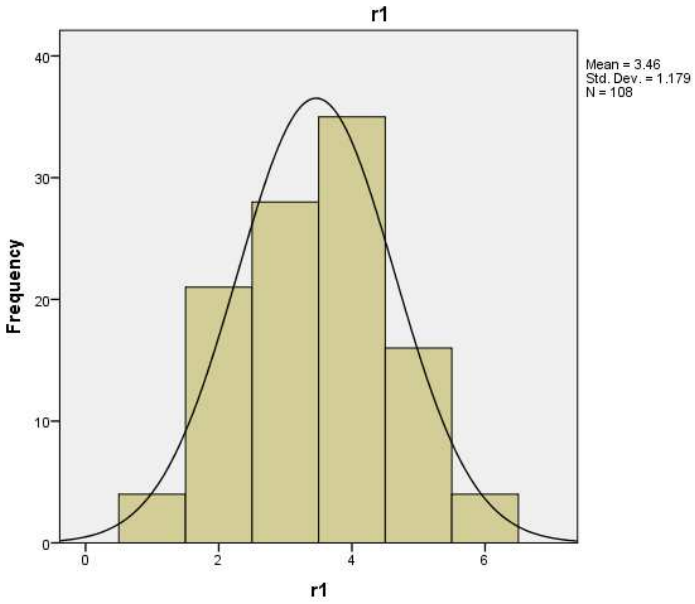
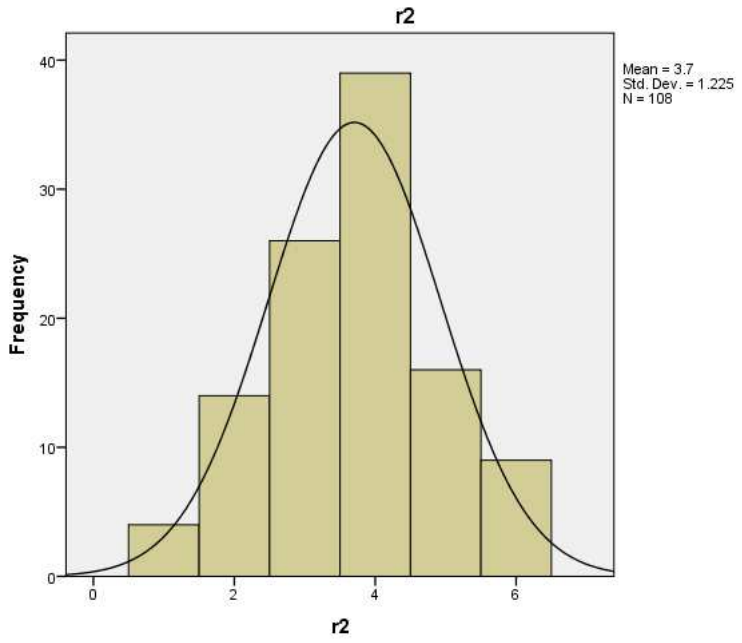
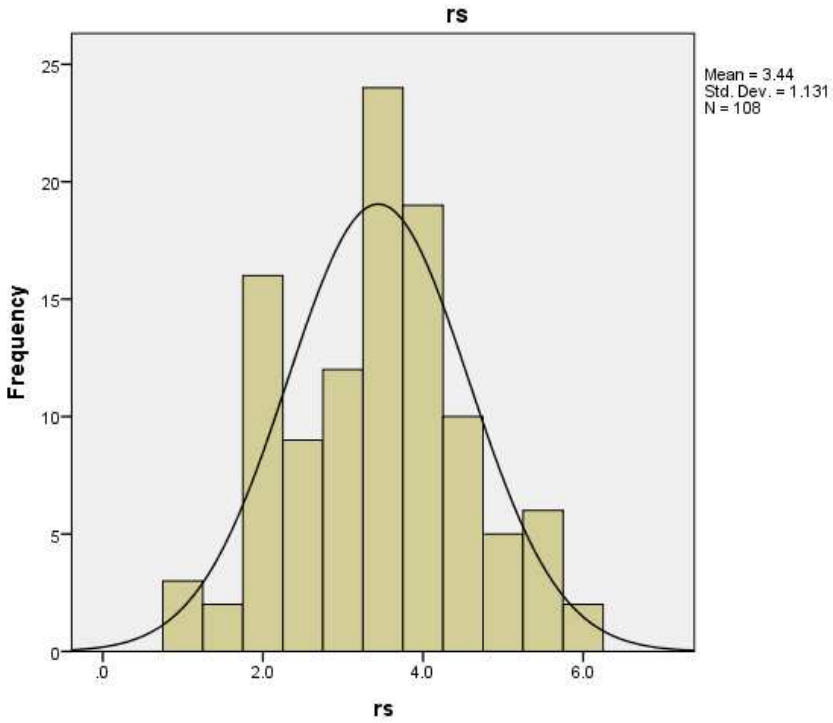


Figure 8. Frequency distribution by score value




r2
Mean = 3.7
Std. Dev. = 1.225
N = 108
Frequency
r2


rs
Mean = 3.44
Std. Dev. = 1.131
N = 108
Frequency
rs

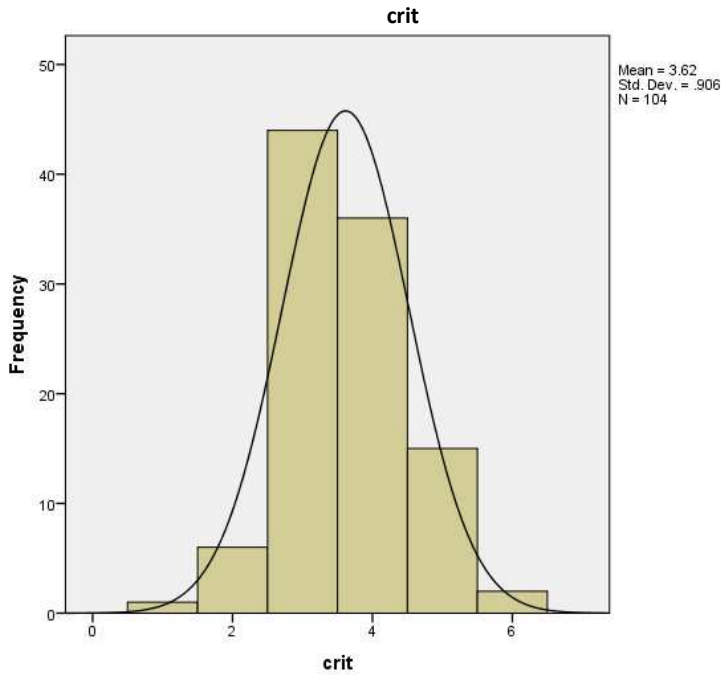


Figure 9. Score frequency distribution by reader (R1, R2, RS, and Crit, respectively).

So, even though our previous measures of consensus (agreement) and consistency indicate that RS and Crit move together more closely than R1 and R2, when frequency distributions are considered, the three human sets of ratings have greater standard deviations, are more widely distributed and make more use of 1 and 6 scores than does *Criterion*, whose SD was less than one, and whose distribution tended toward the mean. The discrepancy helps to underscore the somewhat arbitrary nature of the Turing test that Ellis Page reified—a test involving correlation matrices like the one in Figure 1 from the first chapter. One interpretation of this graph is that fewer essays warranted very low or very high scores on the common rubric than on the college rubric that the instructors used. This interpretation, then, suggests that the common rubric is not as useful for predicting or serving as a basis for grading college essays, because a smaller range of mechanical properties can still generate diversity in rhetorical impact.

Probing further along the lines of distributions brings us to the third notable discrepancy between human and machine scoring: score variance. *Criterion*'s scores were more mean-concentrated than either human's and the Resolved Score. The following graphic (Figure 10) shows the standard deviation and sample variance for each rater. The variance is equal to the average of the squared difference of each score from the mean score. The standard deviation is thought of as the square root of the variance and is expressed in the same units of measurement as the values. Both are measures of dispersion or the spread of values around the mean. All human distributions had positive skew; *e-rater*'s was negative. The computer distribution had higher kurtosis indicating what we see, that the central values are more likely (see Table 3).

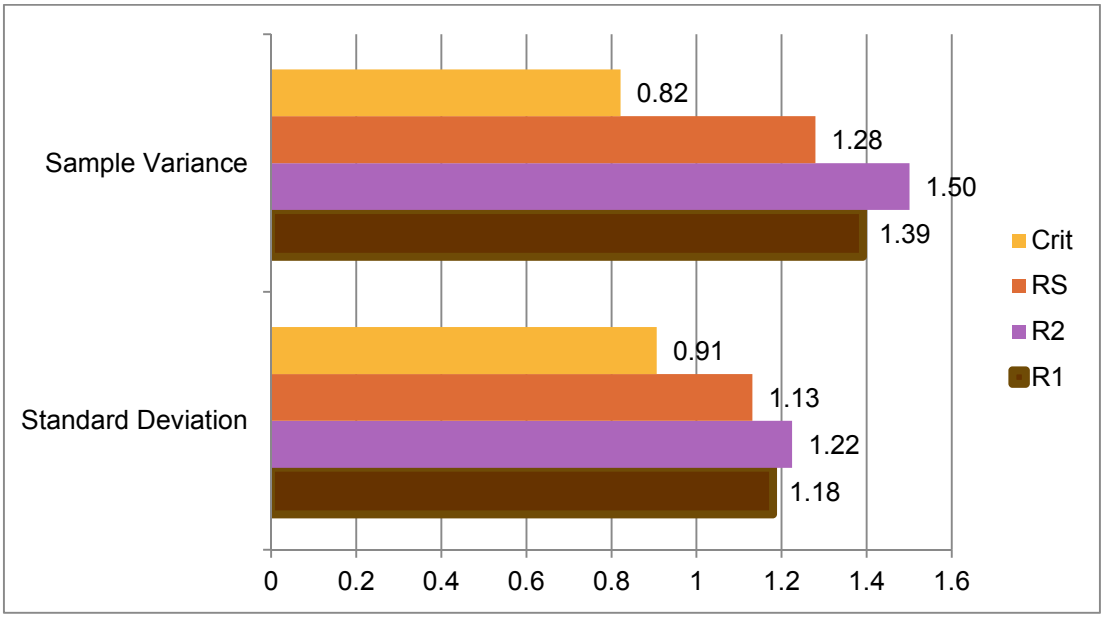


Figure 10. Standard deviations and sample variances for all readers.

Of course, it might be argued that the differences in descriptive statistical measures are due to the separate rubrics used—the fact that humans used the Framework

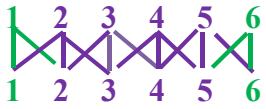
rubric and *e-rater* the common rubric. The most conservative interpretation for the smaller SD for *Criterion* would be that it shows a slight inverse relationship between *Framework* qualities and common qualities. That interpretation would suggest that *Criterion* performs worse in matching *Framework* characteristics for papers that are worse than average (judging them more favorably) and better than average (judging them more poorly).

However, McCurry (2010) also found lower standard deviations and evidence of a central tendency bias in automated scoring as compared to human scoring on a standardized test in Australia, so this finding is not unique to the current dataset. Recall from chapter 3 that standard deviations were .79 and 1.18 for the two software packages on the nine-point scale compared to 1.37 for the human readers. In McCurry (2010)'s study, machines were directly trained on a subset of scores for that test administration, effectively eliminating the possibility that the central tendency finding reflected underlying differences in the construct being evaluated.

There is evidence that *Criterion*'s central tendency is indeed a bias or a construct-irrelevant finding. It is, in fact, thought to be due to the nature of the linear regression algorithm that *Criterion* uses, not due to the fact that, say, the essays read were more homogenous in terms of mechanical errors than in their higher-level qualities. However, in the Shermis and Hamner (2012) study described in chapter 3, the *e-rater* scoring engine did not appear to yield standard deviations drastically different from the resolved score values of the essay sets studied. For the essay set most like the present set, which used a 1-6 scale, the SD of the resolved score was .77; for *e-rater*, it was .79.

Even if the central tendency were not an artifact of the regression calculation as Shermis and Hamner (2012)'s data might suggest, there is another possible construct-irrelevant explanation, based in the personal incentives for readers themselves to maximize agreement. Maximizing agreement implies avoiding the range extremes. Assume two raters are trying to maximize exact agreement. Each rater has a 1 in 6 probability of assigning exactly the same score as the other reader, solely by chance. This can be seen by considering that there are 6^2 possible outcomes of the two score assignments. Six of these possible outcomes, or 6/36, will be agreements (both ones, both twos, both threes, etc.).

However, if readers are trying to maximize *adjacent* rather than exact agreement, as they do in many assessments, then the situation changes. For the numbers at the end of the scale are only adjacent to the sole number next to them, whereas the numbers in the middle are adjacent to two other values. This can be represented diagrammatically in the illustration below:



Selecting a 2, 3, 4, or 5 puts a reader in adjacent connection with three values, or fully half of the 6-point rubric. Selecting a 1 or a 6 puts the reader in connection with only two values or 1/3 of the rubric. Many readers who are under pressure to maximize adjacent agreement—rather than simply to evaluate the particular essay with respect to the features of the rubric given—are going to realize that they have a 50% chance of getting to adjacency just by leaving out 1 and 6 values for any ratings that they assign. (Of course, actual scoring regimes may have protections built in against this plan; for

instance, the CLA (Collegiate Learning Assessment) consists of three analytic, 1-6 scales. Instead of requiring raters to only achieve adjacent agreement with target scores, the CLA requires two adjacent ratings and one exact rating per essay read. The requirement of exact agreement for the one rating means that raters who "play to the center," as it were, stand to fail trials where the scores are straight 1s or straight 6s. A central tendency bias has been well documented in the literature, particularly when raters know they are being monitored and strive to "play it safe" (Leckie & Baird, 2011).

The intuition that scores will be distributed normally, somewhere near the center score value (somewhere between 3 and 4), also explains how human scores patterns can become quite centered around the ratings of central tendency. If a human is choosing between a 2 score and a 3 score, just a basic knowledge of expected probability would portray the 3 as more likely. To quantify this, assume that the mean score of a sample of essays is 3.0 and the standard deviation is 1. According to the empirical rule for normally distributed data, 68% of the responses will be between 2.0 and 4.0, so based on prior probability alone a reader can assume a greater likelihood of attaining the desired agreement by assigning a score within one point of the expected mean. In the case of the present study, 35/108 of the R1 ratings were 4 as were 39 of the R2 ratings—32 and 36 percent respectively. A reader assuming all six ratings (1-6) to be equiprobable would assign a one-in-six or 16.67% prior probability for each rating, but a reader aware of the typical modal rating would realize that any given essay has a roughly 1/3 chance of being a 4 than anything else. Once training sets become mean centered, automated models can take on this characteristic. Whether humans are motivated by probability distributions is unknown; however, personal experience suggests it is quite possible.

One might ask about the interaction between length and mean-centeredness. If *e-rater* is strongly length-associated, and the essay length values are not highly dispersed, could a narrowly-concentrated length distribution explain the concentrated spread of the machine scores? In fact, the 250-300 word essay was the most frequently-appearing one (see Figure 11). The next chart (Figure 12) graphs each of the 30 essays and the scores assigned to them by each reader. Every essay except one received either a 3 or a 4 from *Criterion* whereas R1 ratings ranged from 2-5 and R2 ratings ranged from 2-6. That is to say, the human readers discerned greater variability in quality among essays of roughly the same length.

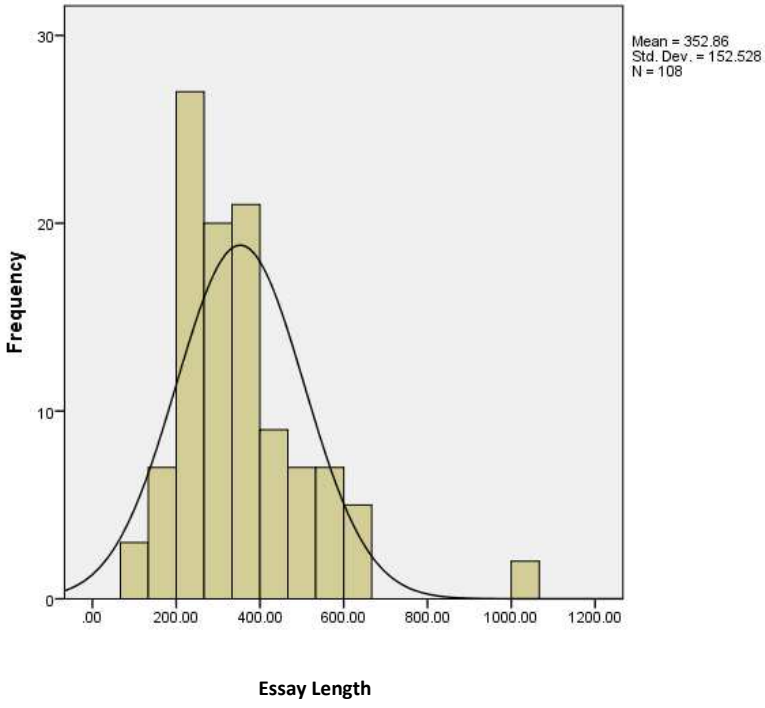


Figure 11. Distribution of essay length in number of words.

As further indication that a highly concentrated length distribution was not fully explanatory of *Criterion*'s concentrated score distribution, Figure 13 shows average essay

length for each of the six possible ratings, for each reader. It is not simply that there were too many average-length essays; *Criterion* tended to award 1s and 6s to essays at only the extremes of the word count range. The steeper slope of the *Criterion* scores provides evidence that *Criterion* used the entire rubric less frequently than human raters even for essays that were relatively shorter or longer. In other words, one had to write an extremely long paper to earn a 6 from *Criterion* or an extremely short paper to earn a 1.

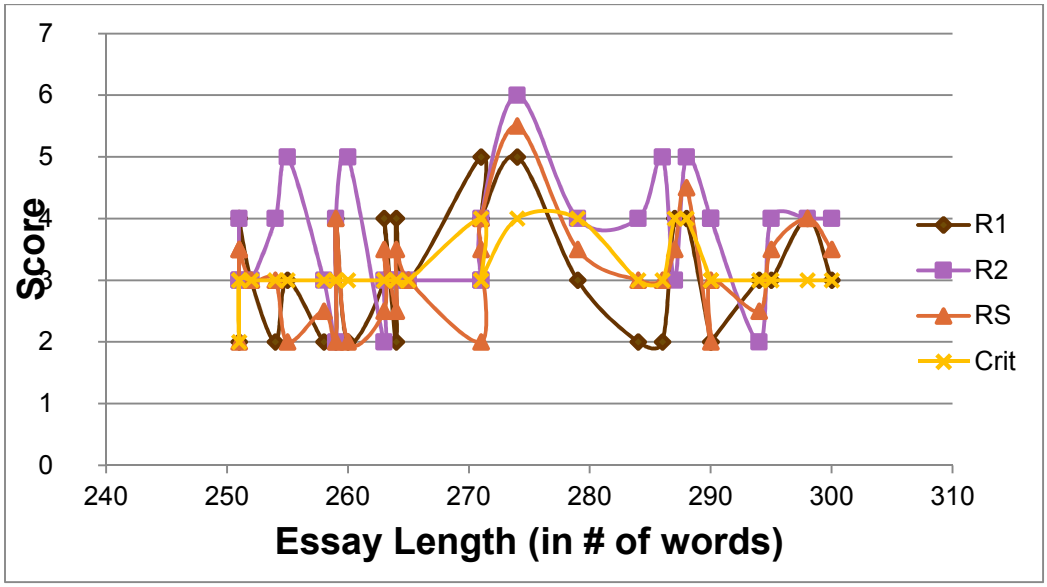


Figure 12. Scores by reader for essays between 250-300 words.

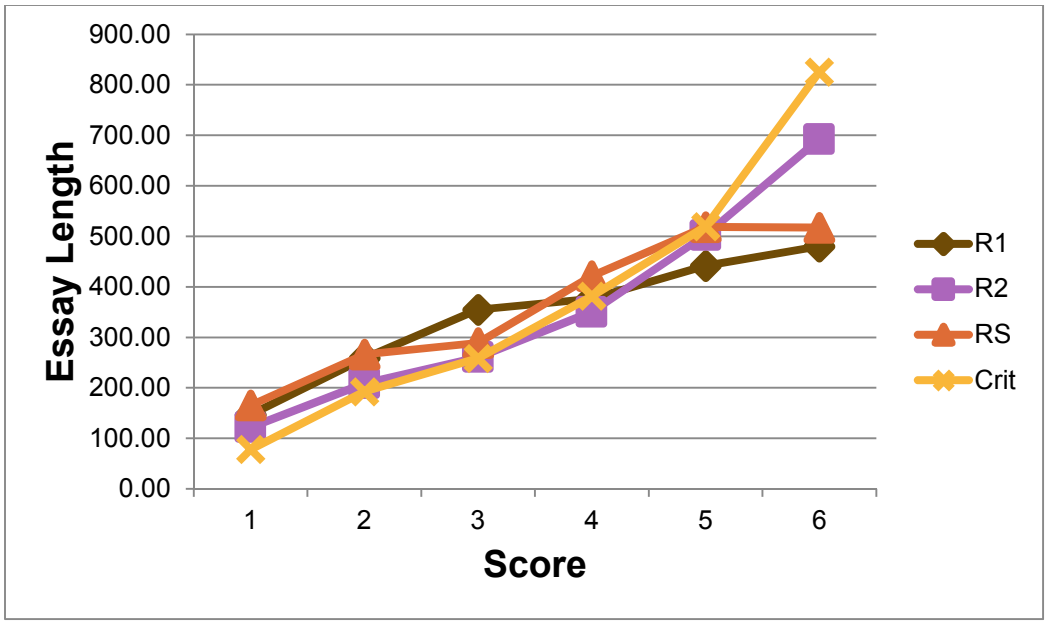


Figure 13. Average essay length per score, by reader

To synthesize the study findings and their relevance: consensus and consistency measures suggest that *Criterion* with its common rubric is just as good at matching reader scores as the readers themselves were, in their first pass at using the Framework-derived rubric. But in terms of distributional differences, what seems to be happening is *Criterion* assigns more mean-centered scores, and this action helps account for the stronger correlations with both readers. *Criterion* tended to "split the difference" when readers disagreed. In keeping with the ratio for the overall sample, *Criterion* assigned scores of 3 or 4 to 75 percent of the split scores (18 of 24 splits). *Criterion*'s scoring also more closely aligns with overall word count.

Toward an Error Analysis

Thus far, the analysis has covered measures of agreement, correspondence of score and essay length, and distributional differences such as mean and variance. What was discovered was that *Criterion* and the human raters performed similarly in terms of agreement measures. On the other hand, human scores were less strongly-associated with length in the human judges than *Criterion* scores, and *Criterion* scores were more closely distributed around the mean. These findings support the "framing" argument made at the outset of this monograph: AEE's success is relative to the analysis tools used to describe it.

Error analysis is yet another tool to frame discourse about AEE, and it can help reveal what might happen to individual students, or to the group of students whose essays did not happen to be assigned exact or adjacent scores by the two readings. A full error analysis lies outside the scope of this chapter; however, the approach will be sketched out here. Assume, that *e-rater* were to agree with human raters 90% of the time. That is a high but not unheard-of exact agreement rate. It still means that ten of every 100 students stand to have their papers mis-scored, at least if the computer is the only rater. Simply talking of disagreement rather than agreement foregrounds those students whose essays stand to be mistakenly handled unless further adjudication were done.

In the case of this study, human raters are the only ones scoring directly on the Framework genre-specific rubric. So their scores have to be held as the standard, even if they are not perfectly reliable. The human raters achieved exact/adjacent agreement in 84 of the cases. In other words, if R1 assigned a 4, R2 assigned a 5 and *e-rater* assigned a 3 the resolved score would be a 4.5 and *e-rater* would be more than a point away from the RS. If both readers assigned the same score and *e-rater* differed by one point, it might

reasonably be considered an error as well because it is a point away from the true score. (See discussions of classical test theory in chapters 3 and 5.) In those 84 non-adjudicated cases, *Criterion*'s *e-rater* scored one point or more lower than the resolved, non-adjudicated score in 12 cases, or 11% of the sample. (The above example would be such a case.) The *e-rater* engine scored one or more point higher than the resolved score in ten or 9.2% of the cases. So, leaving aside the split score cases that had to be adjudicated, having *e-rater* score the non-adjudicated essays by itself would result in an error rate of about 20%. A reasonable conclusion would be that 12 writers would be scored too low and, if this were in placement conditions, possibly would be placed in the incorrect course. In five cases, *e-rater* was off from the RS by more than one point, or six percent of the sample.

It is true that this study is limited by lower-than-ideal human reliability and required 24 adjudications. One could also argue that this just means that essays should always be double-scored. I would support this conclusion. I focus on the accuracy of machine scoring because institutions considering machines for automated placement or rapid assessment may not think they have to institute double-scoring, despite these findings and despite even what best practices might be recommended by ETS and other software developers themselves. Recall that common core administrators had hoped to use AEE alone in scoring essays.

A framework of error reporting exposes the importance of rhetorical framing to assessments of AEE competence. Within a genre-specific rubric, human scores are fundamental and AEE is restricted in its ability to capture scores along this rubric.

Institutions considering using computers for placement without human raters working alongside them stand to mis-score significant percentages of writers.

Moreover, the differences in score distributions together with the analysis of error point toward starkly different approaches that humans and machines have in faced with the task of scoring an essay. Human readers, at least when not constricted by monitoring nor a formalistic rubric, use the extremes of the scoring range more often, and are unwilling to let word count play an unduly important role in determining score. Machines, at least in the case of *e-rater*, cluster their scores more closely around the mean and seem to see overall word count as an indicator of quality.

On the other hand, as seen in chapter 3, *Criterion*'s developers have stated an intention to have the system evaluate only the "same kinds of features" as humans evaluate. This statement is apparently inaccurate, because we know that *e-rater* does not evaluate Framework characteristics. Even a human evaluating "development" and a machine evaluating "development" are evaluating different features of writing: humans are evaluating the way in which the author elaborates ideas and the computers are evaluating the length of discourse elements.

More fundamentally, the current study sheds light on other differences inherent in the scoring process: not only do people and software evaluate different features of writing, people appear to use the rubric more fully and take length less seriously as a determinant of score. In this study's case, the algorithm appeared able to exploit central tendency to promote higher correlations with both readers.

What happens when *e-rater* and a human rater disagree—which score do we "trust"? Because it appears that AEE may be exploiting heuristics such as essay length to

generate its scores, we should apparently not default to the automated software. To study these issues further, an error analysis could be performed on the essay corpus, a qualitative review of each essay and how it was scored, with the hope that certain patterns of rater behavior could be ascertained. Such an analysis is outside the scope of the present study but certainly would be useful.

5. FUTURE DIRECTIONS: ETHICAL AND CRITICAL CONTEXTS

The preceding chapters have pressed toward the following point: AEE's utility is relative to the contexts selected—to the writing prompts chosen, the scoring rubrics used, and the statistical measures applied to the scores generated. When, in addition to statistical patterns, individual cases are considered in evaluating AEE systems, they begin to look quite different from the human raters they have been patterned after. When a rubric is used that is based on the Framework for Success and its Outcomes Statement, computer scoring is likely to mis-handle a significant amount of cases—about 20 percent of the sample, although it does manage to achieve agreement levels similar to the humans in the study.

In this final chapter, I raise broader questions about both machine and human abilities to score essays, even in the context of "common rubrics" that seem to achieve high machine *and* human reliability. Even when a common rubric of chapter of formal writing qualities is employed, use of computer scoring can be problematic. But again, human scoring has been shown to be subject to biases as well. Employing a thought experiment using error analysis (introduced in chapter 4 as a complementary measure of AEE performance) helps flesh out this suspicion. I end, however, on something of a positive tone about the use of computer essay evaluation in global contexts where human pedagogical intervention is sparse.

Let us look first at computer scoring challenges. We have established that scoring essays is not, at this point, an activity with a computational implementation. To review, computation may be framed as the process of expressing a problem such that it can be

calculated by an effective procedure. It is an attempt to describe a problem in step-by-step terms that require no insight or guesswork, with the intent of mechanizing the problem so that it is solvable by machine. In a certain sense, computation is a matter of representation. In an ideal case of representation, a given problem will have a Turing-computable resolution, which is what happens (says the Church-Turing Thesis) to a certain class of problems specifiable in a certain way. When a problem is Turing-computable, the identical problem is represented and performed mechanically (e.g., finding the square root of all numbers between 1 and 25).

To take a biologically-related example of a computational explanation, let us look at the case of vision. In the 1970s and 80s, David Marr (1982) led an effort to understand biological vision systems computationally. His computational theory of vision sought to explain observed biological data (an example of a biologically-sensitive version of computationalism). In Marr (1982), there are three primary levels of analysis of biological systems: the computational, the algorithmic, and the implementation. Marr's insight was to seek to describe high-level behavior in terms of the computation that the system was tasked with. For vision, the eye has to tackle the problem of transforming two-dimensional images of light impinging on the retina into representations of three-dimensional objects. To make stereoscopic vision possible—i.e., three-dimensional vision derived from two receptors located in proximity (e.g., human eyes)—the brain takes account of the disparity in the visual images it receives from each eye and uses that to predict how far away an object is. The problem for a vision system is to select a location on one of the images that can then be identified from the other image so that depth can be computed. After describing the high-level computation to be performed,

Marr and Poggio (1979) proposed a five-step algorithm that would account for how stereoscopic vision overcomes the above problem. Here again, Marr and Poggio's emphasis was on determining the steps that the system uses in order to accomplish the given task, rather than on accomplishing that task without regard to what algorithm might best fit experimental findings about the system being studied. In this work, visual processing is reduced to computations.

But when a problem is not easily represented by an unassailably computational set of procedures, which of course has been the case for most problems involving natural languages and communication, the original problem has to be replaced by a *different* problem before it can then be represented by a set of rules and coded into software. When Ellis Page proposed his set of proxy features, he essentially replaced a set of linguistic qualities that humans recognize (apt word choice, quality of ideas, mechanics, etc…) with a set of qualities that could be coded into machine language, such as length of essay in words, number of paragraphs, number of common words, and so on (1968, p. 216). He replaced the problem of grading an essay with a problem of counting various syntactical features in the essay, and then predicting scores based on that feature set.

Under these circumstances, where the original task or behavior is not itself being expressed in computational terms but rather a new one is being charted, the relationship between the original task and its new redescription needs to be justified. It might help to discuss examples of how this validation occurs. One approach is to simply admit that the new model has limitations as far as its ability to replicate the domain in question. When the system in question is the brain—such as in cognitive science or computational neuroscience—relevant properties of brains can be theoretically represented by

computational models. For instance, in Kintsch's (1998) networked representation of knowledge, each node is thought to be a proposition; the propositions are more or less closely associated with one another (represented by distance in the visual depiction). The drawback to this model is that there is no way, admits Kintsch, to code into the network the relevant associations, since those are acquired by "liv[ing] a normal human life" (p. 417). So we have a representation that has certain limits. But still, the model attempts to model human knowledge construction by providing a possible mechanism—a plausibility case—for how it might be implemented.

Kintsch (1998)'s approach is what Ellis Page (1968) would have called "process" simulation because it works by attempting to emulate the processes by which knowledge is represented in the brain. Its opposite is "product" simulation, or one of modeling outcomes only (say, performing to a certain standard on a recall test). Page stated that product simulation was the only possible method for representing human activity; writing that "*all* computer simulation of human behavior appears to be product simulation rather than process simulation" (p. 214, italics his). Despite his assertion, however, Page also tended to speak as if natural language were actually Turing-computable (as shown in chapter 2 of this monograph) which would imply process simulation at the very least. And in fact, Turing himself seemed to equivocate on whether the Turing test, which is after all a behavioral assessment, required Turing-computability or not. But what Page meant to accomplish in this particular passage was to convey that it was not possible to directly model the values that influenced the decisions humans made about essay quality—just the outcomes of those decisions, i.e., the ratings.

Today, however, it would seem that process simulation, such as what Kintsch (1998) described, is quite useful in the cognitive sciences. Page's claim—simulations can only reproduce outputs, not processes—had roots in the kind of Skinnerian behaviorism that was seen as outdated and replaced by a focus on internal states by scholars such as computational linguist Noam Chomsky, who first critiqued Skinner in 1959, as Yarden Katz has noted (2012, para. 3). "The emphasis [of the Chomskian approach]," Katz wrote, "is on the internal structure of the system that enables it to perform a task, rather than on external association between past behavior of the system and the environment" (2012, para. 6).

Current automated evaluation software continues, however to follow in the Page's "product" simulation mode as developers, often explicitly, do not consider the processes by which humans read and form judgments about essays. Such a behaviorist approach to computer simulation theorizes that what matters is that the computer attain the correct outputs for the case in point, regardless of what computational process it uses to get there. While this approach might seem dated, especially as it traces to Page and B. F. Skinner before him, it is in fact re-emergent. As Katz (2012) went on to note, a neo-behaviorism version of AI research has emerged taking advantage of huge computational power to approach problems with statistical techniques over huge amounts of data, discussion of which is outside the scope of this chapter but which is covered accessibly in Katz (2012).

So, to take stock of the argument: computationally replicating scoring is simply out of reach at this stage, and may be out of reach permanently. Given this, there is a debate in the field of computer science over what forms of simulation are preferable, ones that seek to model and explain internal processes of systems, or ones that seek only to

reproduce their outputs. It appears that AEE software seeks to reproduce outputs only. All of the computer programs that currently score essays are doing so by what Page (1968) calls "product simulation." None of them are either computationally replicating human scoring practices or even attempting to simulate those practices. Instead, they are scoring by proxy measures—none of them are actually measuring the features of the rubric that humans are measuring, even when computers are using the common rubric that was originally described by Diederich and taken up by ETS. (The concept of proxy variables was introduced in chapter 1 of this monograph and described further in chapter 2.)

If computers are product-simulating, this helps account for why AEE systems make mistakes quite different from those of their human counterparts: the mechanisms, speaking at the algorithmic level, not simply the level of physical hardware, driving human and machine raters are quite divergent. People and computers do quite different things when they are rating.

When pairing a process, such as essay scoring, that is not computational and requires fair amounts of human judgement—in other words, people who are not following an algorithm but rather a heuristic guide such as a rubric—with computers that are following algorithms designed to human judgement, there is likely to be error. Here I advance a partial argument that such error would be introduced in a common rubric situation, not only in the college-specific rubric studied in chapter 4. Examining parts-of-speech tagging helps further explicate this concern, because even at lower levels of analysis, natural language processing (NLP) seems to be the kind of system not composed of Turing-computable rules.

According to Manning (2011) humans tasked with annotating the parts of speech on any given corpus achieve about 97% agreement on parts of speech tagging tasks, and machines can match that accuracy as well. A parts-of-speech tagger is learning algorithm, not unlike essay rating software, that analyzes a pre-tagged corpora of text, using probability metrics to predict the parts of speech correctly, and then applies its weighted algorithm to new text. But when Manning (2011) analyzed the errors that each made, he found that some of the human errors appeared to be cases of carelessness or inattention to tags that should have been unambiguous whereas the machines, while "quite good, regularly make egregious errors" (p. 2). For instance, a speech tagger unfamiliar with the word "substandard" tagged it as a noun rather than an adjective (p. 6). In other cases, Manning found that there was no clear cut choice for assigning a tag. What becomes clear through a foray into parts-of-speech tagging is that not even at the pre-syntactic level does the problem of accuracy resolve itself. Whether simulating score *production* or the *process* of score generation, error will arise; machines and humans will err differently.

This evidence of error in even the lowest stages of natural language processing is the proverbial tip of the iceberg in suggesting that natural language presents problems for computer simulations at various levels of linguistic analysis. As a result, computers will make mistakes that people would find silly—whether awarding high marks to long essays full of gibberish or making bizarre parts-of-speech misclassifications. So it is not just that computer scoring cannot capture the construct of a genre-specific rubric such as the Framework rubric. Even along the common rubric, computer simulations will err.

Here, then, is the problem I referred to at the outset of this chapter: if computer scoring is flawed in this sort of definable way, even at high levels of interrater agreement,

mistakes are bound to happen. In any scoring session, we might expect a percentage, even small, of the sort of "egregious" scoring errors that Manning (2011) calls out. Such scoring errors are precisely what Les Perelman has identified with his babel generator (introduced in chapter 2). These errors, however, will not be obvious simply by looking at a series of scores. The errors could only be identified by an examination of each essay, similar to the process initiated in chapter 4 here and what Manning (2011) did in his error analysis. But an analysis of essays and their nature involves independently assessing essays, in other words, determining where they fall on the score continuum. In other words, if we are worried about the kinds of mistakes that Manning (2011) identified but at the level of overall quality judgement we would have to review each essay and essentially decide whether the computer were correct about the score, or not. We would have to score the essay, that is to say. *Therefore, computer scoring does not alone serve as a legitimate method of scoring.*

On the other hand, computer scoring might be part of a legitimate system of scoring, by serving as the second or "read behind" reader set to check whether human scores are on target. The next section works on a more precise understanding of what it means for human scores to be "on target" and for computer scoring to check for that. Here is the intuitive argument: if it were possible to derive a human "error" rate, as well as a rate at which the computer could be clearly off the mark as described above, having both score essays would reduce error. Assume, just for the sake of argument, that there were such thing as both a human error rate and a machine error rate. Divergent scores would call attention to a mistake on either side and signal an adjudication. Assume just for the sake of the Gedanken experiment that both error rates are .3. That is to say, each

makes a mistake on three of 100 essays. The probability that both will err on the same essay is .03 x .03 = .0009. The probability that both would err on the *same* essay by assigning it the *same* score, meaning that the error would go unnoticed, is .0009 x (6/36) = .00015. In other words, there is less than a one in ten thousand chance that the computer and the person will erroneously issue the same score—an undetectable mistake. In principle, at least, there is a case for computers to score alongside people, calling attention to errant scoring.

But some might argue that the notion of a human error rate is equivocal. This issue will not be fully resolved here, but can at least be briefly addressed.

Human Error in Assessment

Something the educational measurement, writing assessment, and AEE developers will agree on is that human raters may legitimately disagree as to their judgments of essay quality. Essay assessment has no "gold standard" for measuring performance in the way that a shooting competition or a road race might have. In those more defined contexts, what counts as a bullseye or a sub five-minute mile is clearly delimited. But as Shermis and Hamner (2012) noted, "Trained human raters, as with subject-matter or writing experts, can read the same paper and assign the same or different scores for different reasons" (p. 19). The idea here is that even expert readers may disagree, without one or the other being obviously wrong. For that reason, deriving a definition of human rater error is a puzzle. Two raters could disagree and still assign valid scores.

The more serious problem, however, is that human raters can not only legitimately disagree, but can fall victim to rater biases as well. This presents an

epistemological problem: there is no *a priori* way to tell whether discrepant scores are the result of legitimate disagreement—or of bias—just by looking at them. While a certain percentage of computer scores may be clearly erroneous as discussed above, human mistakes are likely to be more subtle. As we discussed in the preceding section, having multiple raters reduces the chances that both are in error, particularly when they arrive at the same score, but does not eliminate the problem entirely.

Classical test theory, according to some, holds any discrepant scoring as error. Weigle (1998), citing Linacre (1989), wrote that "true-score [i.e., classical test] approaches to the problem of rater variation see such variation as undesirable error variance, which must be eliminated or reduced as much as possible…" (Weigle, 1998, p. 264). That is to say, classical test theory would have all human discrepancies classified as error. True score theory is, thus, a realist doctrine, in that there is a fact of the matter concerning an essay score relative to a rubric, and each rating represents an approximation of that score (the observed value plus error).

But Les Perelman (2013) offered a slightly distinct interpretation of classical test theory, suggesting a role for adjacency as non-errant scoring, because "adjacent agreement in the correct direction between two readers (e. g. one rater gives an essay a score of 3 and the second rater gives the essay a score of 4) will more closely approximate a True Score of 3.4 than two scores of 3" (2013, p. 6). For Perelman, adjacent agreement is assumed to contribute to accuracy. Assigning more than one rater, the common practice, makes sense on this model because multiple ratings will, according to the theory, approximate the limit of the true score. Assigning more than one rater also reduces the probability that both scores will be due to human rater bias.

Deriving a theory of human error will depend on the philosophy of measurement adopted. As noted, Weigle (1998), for instance, documented the true score approach that sees rater variation as merely undesirable error (p. 264); the Rasch model that advocates training raters to be self-consistent rather than in agreement with each other (p. 264); and a third empirically-driven approach that asks whether rater training can actually create inter-rater consistency or rather simply creates self-consistency among raters. A common thread in these differing views is that raters will inevitably disagree. Sometimes disagreement is problematic; other times it is acceptable.

We might further say that human scores are not in error if they result from raters who agree, to borrow Weigle's language, "on a definition of the construct being measured in an essay examination" (1998, p. 281). In these and only these cases, then, let us consider scores to be in reasonable agreement or disagreement with one another. The term "reasonable" is inspired by a related idea honed in political philosophy, where idealizations are often made that involve individual citizens possessing certain attitudes and capacities. For example, political philosopher John Rawls wrote that societies should be characterized by "reasonable pluralism" (Rawls & Kelly, 2001, p. 3). Roughly speaking, for Rawls, a reasonable sort of pluralism is predicated upon citizens' honoring what are agreed to be fair terms of cooperation with one another. While one citizen practices Wicca and another Hinduism, say, their pluralist religious views are *reasonable* when they agree to honor each other's religious holidays at their workplace. Reasonability carries with it a "moral sensibility" (Rawls & Kelly, 2001, p. 7). Reasonable agreement or disagreement means that the scores were arrived at by a considered comparison of the rubric to the essay at hand.

A reasonable (and non-errant) score discrepancy exists in the rating context when trained raters reasonably disagree over a particular essay, if both arrived at their ratings through an earnest effort to deploy the values of the relevant rubric to that particular performance, and if both ratings could be justified by the raters by appeal to the rubric. What qualifies psychologically as an "earnest" effort by a rater to apply a rubric to an essay leaves some room for interpretation at this stage.

Human raters sometimes do fail the reasonability test. Threats to validity include construct irrelevant sources of variation such as severity drift, halo, and central tendency (Leckie & Baird, 2011, p. 399-401). Some of these reliability threats are addressed through norming and training, but findings differ as to whether they persist nevertheless. Disagreement would *not* be reasonable if, for instance, one reader got tired and marked every essay a "5" because that reader would not be able to offer justifiable reasons for applying the score to the essay. Nor would disagreement be reasonable if one reader decides that what is meant by critical thinking was writing the most flowery sentences possible while the other applied more typical criteria of, say, degree to which author provides supporting arguments. To be sure, our notion of reasonableness is loosely defined, but at least intuitively plausible.

The preceding analysis helps us answer the question from the prior section. Recall that earlier we expressed the worry based on an analysis of parts-of-speech tagging, that computers had the potential to make egregious errors even when tying their scores to a common rubric. We saw that Les Perelman's babel generator reveals the way a computer scoring engine can be fooled by long essays with long, grammatically-correct sentences with unusual vocabularies. Say further that we can derive a percentage of human scores

that would also be in error (i.e., not based in reasonable disagreement). Deriving an average human error rate might involve the percentage of instances of either non-adjacent agreement or adjacent agreement that was determined not to be reasonable because one of the conditions above did not hold. Assuming that we could derive a predicted human error rate from past experience, computer scoring over a common rubric could help detect errors by a human rater: the probability that an undetectable error would occur (by both assigning the same score) would be the human error rate multiplied by the computer error rate multiplied by 1/6. (An example was given earlier.)

Because people make mistakes, and even double-scored essays can mask errors that people make, Lottridge, Schulz and Mitzel (2013) advanced the proposal that AEE be used in a read-behind fashion, to monitor the kinds of human bias effects that have been categorized as "drift." Given that drift can be hard to verify statistically and rule out underlying differences in the essays presented over time and other variables, the authors trained a computer model on a sample of scores and used that system to identify non-adjacency with human reads and flag those for adjudication. While there is still no guarantee that the computer scores are more accurate than the human scores in cases of non-adjacency, it is certainly possible that problematic human scores could be flagged and adjudicated through this process. And as indicated, having more than one reader increases the probability of error detection.

This analysis of when computers and people may be said to err with respect to rating essays is not meant to be an exhaustive catalog. As Manning (2011) described in his error analysis, there are many other types of errors, including those where the correct answer is in fact ambiguous. I suspect that ambiguity may be the largest problem for

essay scoring, because there is no way to know *a priori* what the correct score of any given essay should be. The only way to determine the score is to score the essay. Adjudications over discrepant scores are themselves not guaranteed to be free of construct-irrelevant variance.

Computer Essay Evaluation and Global Poverty

While current automated techniques lack the justification to score essays with no human back-up, there may be roles for formative feedback in the broader composition landscape. The following section sketches—in broad, impressionistic strokes—possible roles for AEE software, including in developing parts of the world where computer technology is easier to come by than English educators.

One such role originates, actually, in a social critique of AEE. There is a concern articulated by Perelman (personal communication, 2014) and other critics of automated essay evaluation in the college setting, including Herrington and Moran (2001) that amounts to a social critique. As the latter authors observed, the problems of large classes and high student/faculty ratios will not disappear if AEE is brought in to solve the challenge of scoring multitudes of essays. Rather, it will serve to insure "that these problems will remain and become structural and permanent" (p. 496). Elite institutions will better be able to afford robust composition programs with the human interaction that we know is important to building better writers, whereas poorer institutions may turn to automation and the divide grow larger.

The specter of computer scoring exacerbating the digital divide—actually, a reverse digital divide where lower status institutions over-rely on technology—points

toward potential challenges of using even better software than that currently available. Even a next-generation system that had more valid scoring and feedback capability would not be able to provide the social interaction that many feel drives writing improvement. One might build a case for a normative right to be evaluated by a being with the ability to empathize, or at least, the right to make choice.

On the other hand, it is not unthinkable that the tremulous beginner may prefer being evaluated by a "dumb" machine incapable of ridicule or disdain, than by a crusty grammarian with red pen in hand. Writing to a computer that might be able to provide grammatical or other limited help could conceivably boost students who are reluctant to turn in drafts to teachers. As anyone who has Google-searched a sensitive subject knows, computer-mediated discourse has the advantage of assuring the searcher's anonymity, at least, vis-à-vis the search engine. In fact, research showed that Edward Snowden's leaks resulted in a decrease in searches of topics the NSA might find offensive (Mathews & Tucker, 2014), suggesting that indeed it is worry about the *humans* that might be intercepting our searches, not the search engines themselves, who necessitate our subterfuge.

And while human grading may be preferable to AEE, the globally disadvantaged may not have that choice: they may be choosing between technology-mediated education or little to no education at all. Far from increasing economic divides, automated evaluation could be seen as just another technology reducing the education gap, of a piece with online learning platforms such as the well-known Khan Academy, which provides online tutorials and exercises in a variety of subjects. Those not having regular access to teachers could benefit from interactive writing technology. So while AEE may

indicate lower-end educational opportunity, it may be the only opportunity available to some.

A final point not to be forgotten is that humans and machines are often pitted in the discourse about automated scoring and evaluation as if the two are locked in a zero-sum game: as if more machines equals fewer people in the classroom. While this may be a legitimate material fear of those against computer pedagogies of all kinds, it does not have to be the overshadowing reality all the time. Machines should supplement human teaching, not replace it.

The Machine/Human Interface

Haraway (1985/2010) wrote that technology is imbricated in power, musing that "writing, power, and technology are old partners in Western stories of the origin of civilization," (p. 280). She also suggested, a bit offhandedly, that it might be those technology workers themselves alongside others (those "unnatural Cyborg women making chips in Asia and spiral dancing in Santa Rita jail" (p. 281)) whose hybridized identities will bring about productive discourse.

Compare what Haraway suggests to the dialectical approach of White and McAllister (2006), who suggest that, "rather than trying to tell the history of computer-assisted writing assessment as a tale of good and evil—where good and evil could be played interchangeably by computers and humans—we prefer to tell the history more dialectically, that is, as a history of interested complicities" (p. 11). For White and McAllister, the plurality of positions one might stake out about automated scoring are each justifiable in some way.

What will happen with AEE? It is not the star wars apocalypse. If my research has shown anything, it has shown that people, and the computers they invented, learn and act in distinct ways. We ought to exploit the benefits of each. Our future calls for a hybridized reality; it calls for one in which people and technologies continue their cyborg work together. Computers, as powerful as they are in sheer computing speed and accuracy, reveal some stark limits in the domain of natural language. These performance limits are, of course, ever-changing in the shifting sands of innovation. Alan Turing and Ellis Page were too optimistic about the computability of language. But the opposite mistake would be to declare computationalism irrelevant to communication. Finding ways to exploit the value added by machines is not to devalue humans; indeed, if done well it can elevate the least advantaged of all.

APPENDICES

Approved informed consent documents from URI and SCSU

Southern Connecticut State University
SCSU SCHOOL OF GRADUATE STUDIES

Date: 12/18/13 Re: **Expedited** Protocol and Consent Approval

Attn: Ms. Catherine Barrett
21 Bellevue Road
New Haven, CT 0651 1

Protocol Title: Machine scoring enters the classroom

Protocol Number: 13-226 Approval period: 12/18/13 – 12/17/14

Department: English

Dear Ms. Barrett,

On behalf of the Southern Connecticut State University Institutional Review Board (IRB), I am pleased to inform you that the protocol listed above, has been reviewed and approved as submitted for the period indicated.

The SCSU HRPP IRB operates under the Code of Federal Regulations CFR: Title 45, part 46. As a result this approval is granted with the understanding of continuing investigator responsibilities. Initiation of the research coverec by this approval will be considered acceptance of the following responsibilities

1. The attached consent form must be used as is, unless a subsequent modification is approved by the IRB (copies may be made). The approved consent form has been stampec with an expiration date and initialed by the IRB chair.
2. If data collection is to continue beyond the expiration date indicated in this letter, and stamped on your consent form, the IRB must be informed using the ***Continuing Review Form***, **prior** to the expiration date, otherwise, you must cease data collection as your research will no longer he approved (enclosed).
3. Changes in procedures which in any way influence the research participants, study methodology, consent or protocol, must be submitted in writing **in advance** to the IRB for approval using the ***Request for Revision Form*** (enclosed).
4. A final progress report must be submitted to the IRB by the Principal Investigator(s) within 30 days of study termination using the ***Research Completed Form*** (enclosed). Please note: if this research is a thesis or dissertation completed in partial fulfillment of degree requirements, this form must be submitted as part of your final thesis/dissertation submission to the School of Graduate Studies
5. If, during the conduct of your research, any adverse events occur involving the research participants, an ***Adverse Event Form*** must be completed and submitted to the IRB immediately (enclosed).
6. In the completed presentation of your research project, please be sure to maintain all privacy and confidentiality components promised to participants in your consent/assent document(s).

The IRB welcomes your research project into the list of approved protocols. Your compliance with the above conditions will help protect your research for the approval period and permit final allowance of your research activity.

Sincerely,

Dr. W. Jerome Hauselt, IRB Chair
School of Graduate Studies
Voice: 203-392-5243, FAX 203-392-5221
Email: hauselw1@southernct.edu

Enclosures (Revised 07/27/09)

Research Protection
Southern Connecticut State University

501 Crescent Street • New Haven, Connecticut 06515-1355 • (203) 392-5240 • fax (203) 392-5276 • www.SouthernCT.edu/grad
AN INSTITUTION OF THE CONNECTICUT STATE UNIVERSITY SYSTEM • AN EQUAL OPPORTUNITY UNIVERSITY

THE
UNIVERSITY
OF RHODE ISLAND
DIVISION OF RESEARCH
AND ECONOMIC
DEVELOPMENT

OFFICE OF RESEARCH COMPLIANCE
70 Lower College Road, Suite 2, Kingston, RI 02881 USA
p: 401.874.4328 f: 401.874.4814 uri.edu/research/tro/compliance

THINK BIG WE DO℠

DATE:	January 27, 2014
TO:	Robert Schwegler, PhD
FROM:	University of Rhode Island IRB
STUDY TITLE:	[525638-2] Automated Essay Evaluation and the Computational Paradigm: Machine Scoring Enters the Classroom
IRB REFERENCE #:	HU1314-071
SUBMISSION TYPE:	Revision
ACTION:	APPROVED
APPROVAL DATE:	January 18, 2014
EXPIRATION DATE:	January 17, 2015
REVIEW TYPE:	Expedited Review
REVIEW CATEGORY:	Expedited review category # 7

Thank you for your submission of Revision materials for this research study. University of Rhode Island IRB has APPROVED your submission. This approval is based on an appropriate risk/benefit ratio and a study design wherein the risks have been minimized. All research must be conducted in accordance with this approved submission.

This submission has received Expedited Review based on the applicable federal regulation.

Please note that any revision to previously approved materials must be approved by this office prior to initiation. Please use the appropriate revision forms for this procedure.

All SERIOUS and UNEXPECTED adverse events must be reported to this office. Please use the appropriate adverse event forms for this procedure. All FDA and sponsor reporting requirements should also be followed.

Please report all NON-COMPLIANCE issues or COMPLAINTS regarding this study to this office.

Please note that all research records must be retained for a minimum of three years.

Based on the risks, this project requires Continuing Review by this office by January 17, 2015. Please use the appropriate renewal forms for this procedure.

If you have any questions, please contact us by email at compliance@ds.uri.edu. Please include your study title and reference number in all correspondence with this office.

Please remember that informed consent is a process beginning with a description of the study and insurance of participant understanding followed by a signed consent form. Informed consent must continue throughout the study via a dialogue between the researcher and research participant. Federal regulations require each participant receive a copy of the signed consent document unless the signature requirement has been waived by the IRB.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

Anson, C. M., Dannels, D. P., Flash, P., & Housley Gaffney, A. L. (2012). Big rubrics and weird genres: The futility of using generic assessment tools across diverse instructional contexts. *Journal of Writing Assessment, 5*(1). Retrieved from http://www.journalofwritingassessment.org/article.php?article=57

Attali, Y. (2007). *On-the-fly customization of automated essay scoring.* (ETS Research Report No. RR-07-42). Princeton, NJ: Educational Testing Service.

Attali, Y. (2013). Validity and reliability of automated essay scoring. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 181-198). New York: Routledge.

Baker, N. L. (2014). Get it off my stack: Teachers' tools for grading papers. *Assessing Writing, 19*, 36-50.

Bazerman, C. (1999). *The languages of Edison's light.* Cambridge, MA: MIT Press.

Bazerman, C., Dean, C., Early, J., Lunsford, K., Null, S., Rogers, P., & Stansell, A. (Eds.). (2012). *International advances in writing research: Cultures, places, measures.* Fort Collins, CO: The WAC Clearinghouse and Parlor Press.

Berlin, J. (1987). *Rhetoric and reality: Writing instruction in American colleges, 1900-1985*. Carbondale: Southern Illinois University Press.

Bermúdez, J. L. (2010). *Cognitive science: An introduction to the science of the mind.* Cambridge, U.K.: Cambridge University Press.

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

Broad, B. (2006). More work for teacher? Possible futures of teaching writing in the age of computerized writing assessment. In P. F. Ericsson, & R. H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 221-233). Logan, UT: Utah State University Press.

Broad, B. (2012). Mapping a dialectic with Edward M. White (in four scenes). In N. Elliot, & L. Perelman (Eds.), *Writing assessment in the 21st century: Essays in honor of Edward M. White* (pp. 259-270). New York, NY: Hampton Press.

Burke, K. (1966). *Language as symbolic action*. Berkeley and Los Angeles, CA: University of California Press.

Burstein, J. (2009). Opportunities for natural language processing research in education. *Springer Lecture Notes in Computer Science*, 5449, 6–27.

Burstein, J., & Chodorow, M. (1999). Automated essay scoring for nonnative English speakers. *Proceedings of the Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing.* Joint Symposium of the Association of Computational Linguistics and the International Association of Language Learning Technologies (pp. 68-75). College Park, Maryland.

Burstein, J. C., Chodorow, M., & Leacock, C. (2004). Criterion online essay evaluation: An application for automated evaluation of student essays. *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence,* Acapulco, Mexico. Association for the Advancement of Artificial Intelligence.

Chen, C.-F. (E.) & Cheng, W.-Y. (E.) (2008). Beyond the design of automated writing evaluation: pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology*, *12*, 94-112.

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

Chodorow, M. & Burstein, J. (2004). *Beyond essay length: Evaluating e-rater's performance on TOEFL essays.* (TOEFL Research Report No. RR-73). Princeton, NJ: Educational Testing Service.

College Board. (n.d.). ACCUPLACER frequently asked questions. Retrieved from https://accuplacer.collegeboard.org/professionals/frequently-asked-question#08

Council for Aid to Education. (n.d.). CLA+ technical FAQs. Retrieved from http://cae.org/images/uploads/pdf/CLA_Technical_FAQs.S13.pdf

Coe, M., Hanita, M., Nishioka, V., & Smiley, R. (2011). *An investigation of the impact of the 6+1 trait writing model on grade 5 student writing achievement: Final Report.* Washington, DC: Institute of Education Sciences, U.S. Department of Education.

Colford, P. (2014). A leap forward in quarterly earnings stories. *The definitive source: a look inside the world's most trusted news organization.*

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing*, 18(1), 100-108. doi:10.1016/j.asw.2012.11.001

Condon, W. & Kelly-Riley, D. (2004). Assessing and teaching what we value: The relationship between college-level writing and critical thinking abilities. *Assessing Writing 9*(1), 56-75.

Cizek, G. C. & Page, B. A. (2003). The concept of reliability in the context of automated essay scoring. In M. D. Shermis & J.C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 125-145). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Churchland, P. S., & Sejnowski, T. J. (1994). *The computational brain*. Cambridge: MIT Press.

Copeland, B. J. (2002). "The Church-Turing Thesis." The Stanford Encyclopedia of Philosophy (Fall 2002 Edition). E. N. Zalta, URL = http://plato.stanford.edu/archives/fall2002/entries/church-turing/.

Copeland, B. J. (2004). *The Essential Turing.* Oxford: Oxford UP.

Council of Writing Program Administrators. (2014). WPA Outcomes Statement for First-Year Composition (3.0).

Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences, 108*(17), 6889-6892. doi:10.1073/pnas.1018033108

Deane, P. (2013). On the Relation Between Automated Essay Scoring and Modern Views of the Writing Construct. *Assessing Writing, 18*(1), 7-24.

Deane, P., Williams, F., Weng, V., & Trapani, C.S. (2013). Automated Essay Scoring in innovative assessments of writing from sources. *Journal of Writing Assessment, 6*(1). Retrieved from http://www.journalofwritingassessment.org/article.php?article=65

Diederich, P. B. (1996). Turning Fords into Lincolns: Reminiscences on teaching and assessing writing. *Research in the Teaching of English, 30*(3), 352-360.

Diederich P. B., French J. W., & Carlton S. T. (1961). Factors in judgments of writing ability. *ETS Research Bulletin Series*. 1961(2):i-93. doi:10.1002/j.2333-8504.1961.tb00286.x

Dyehouse, J. (2007). Knowledge consolidation analysis: Toward a methodology for studying the role of argument in technology development. *Written Communication, 24*(2), 111-139.

Edsurge. (2012). The end of multiple choice? The quest to create accurate robot essay graders. Retrieved from www.fastcoexist.com

Ericsson, P. F. (2006). The meaning of meaning: Is a paragraph more than an equation? In P. F. Ericsson, & R. H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 28-37). Logan, UT: Utah State University Press.

Ericsson, P. F., & Haswell, R. H. (Eds.) (2006). *Machine scoring of student essays: Truth and consequences*. Logan, UT: Utah State University Press.

Follman, J. C., & Anderson, J. A. (1967). An investigation of the reliability of five procedures for grading english themes. *Research in the Teaching of English, 1*(2), 190-200. Retrieved from http://www.jstor.org/stable/40170454

Foucault, M. (1975). Panopticism. In D. Bartholomae & A. Petrosky (Eds.), *Ways of reading: An anthology for writers* (pp. 282-309). Boston, MA: Bedford/St. Martin's.

Frankfurt, H. (2004). *On Bullshit*. Princeton, NJ: Princeton University Press.

Galton, A. (2006). The Church–Turing Thesis: Still valid after all these years? *Applied Mathematics and Computation, 178*(1), 93–102. http://dx.doi.org/10.1016/j.amc.2005.09.086

Hagerman, C. (2011). An evaluation of automated writing assessment. *The Jaltcall Journal*, *7*(3), 271–292.

Haraway, D. (2010). A manifesto for cyborgs: Science, technology, and socialist feminism in the 1980s. In V. B. Leitch, W. E. Cain, L. A. Finke, B. E. Johnson, J. McGowan, T. D. Sharpley-Whiting & J. J. Johnson (Eds.), *The Norton Anthology of Theory & Criticism* (2nd Edition) (pp. 2190-2220). New York: WW Norton & Company. (Original work published 1985)

Haswell, R., & Wilson, M. (2013). Professionals against machine scoring of student essays in high-stakes assessment. Retrieved from www.humanreaders.org

Haswell, R., Donnelly, W., Hester, V., O'Neill, P., & Schendel, E. (2012). An Annotated Bibliography of Writing Assessment: Machine Scoring and Evaluation of Essay-length Writing. *Journal of Writing Assessment 5*(1).

Haswell, R. (2005). NCTE/CCCC's recent war on scholarship. *Written Communication* 22.2, 198-223.

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

Herrington, A., & Moran, C. (2006). WritePlacer plus in place: An exploratory case study. In P. F. Ericsson, & R. H. Haswell (Eds.), *Machine scoring of student*

essays: Truth and consequences (pp. 114-129). Logan, UT: Utah State University Press.

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* (pp. 219-232). New York, NY: Hampton Press.

Isaacs, E., & Molloy, S. (2010). Texts of our institutional lives: SATs for writing placement: a critique and counterproposal. *College English*, *72*(5), 518-538.

Jeuniaux, P., et al. (2006). Cognitively inspired NLP-based knowledge representations: further explorations of Latent Semantic Analysis. *International Journal On Artificial Intelligence Tools 15*(6), 1021-1039.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York, NY: Cambridge University Press.

Klobucar, A., Deane, P., Elliot, N., Ramineni, C., Deess, P., & Rudniy, A. (2012). Automated essay scoring and the search for valid writing assessment. In C. Bazerman, C. Dean, J. Early, K. Lunsford, S. Null, P. Rogers, & A. Stansell (Eds.), *International advances in writing research: Cultures, places, measures.* (pp.103-119). Fort Collins, CO: The WAC Clearinghouse and Parlor Press.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2/3), 259-284.

Latour, B. (1999). *Pandora's hope: essays on the reality of science studies*. Cambridge, MA: Harvard University Press.

Lauer, J. (2004). *Invention in Rhetoric and Composition*. West Lafayette, IN: Parlor Press, LLC.

Leckie, G., & Baird, J. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement, 48*(4), 399-418. doi:10.1111/j.1745-3984.2011.00152

Levy, L. B., & Fritz, K. V. (1972). Status report on the computer grading of essays. *Counseling Center Reports, 5*(10).

Leydens, J., & Olds, B. Complicating the fail-or-succeed dichotomy in writing assessment outcomes. In N. Elliot & L. Perelman (Eds.), *Writing assessment in the 21st century: Essays in honor of Edward M. White* (pp. 425-438). New York, NY: Hampton Press.

Lim, H., & Kahng, J. (2012). Review of Criterion. *Language Learning & Technology*, *16*(2), 38–45.

Lottridge, S. M., Schulz, E. M., & Mitzel, H. C. (2013). Using automated scoring to monitor reader performance and detect reader drift in essay scoring. In M. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 233-250). New York: Routledge.

Lyotard, J. (2010). Defining the postmodern. In V. B. Leitch, W.E. Cain, L.A. Finke, B.E. Johnson, J. McGowan, T.D. Sharpley-Whiting & J.J. Johnson (Eds.), *The Norton Anthology of Theory & Criticism* (2nd Edition) (pp. 1465-1468). New York: WW Norton & Company. (Original work published 1986)

Maddox, T. T. (2006). Piloting the COMPASS E-write software at Jackson State Community College. In P. F. Ericsson, & R. H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 147-153). Logan, UT: Utah State University Press.

Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? *CICLing'11 Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing,* 171-189.

Mason L., Tornatora M., & Pluchino P. (2013). Do fourth graders integrate text and picture in processing and learning from an illustrated science text? Evidence from eye-movement patterns. *Computers & Education 60*(1), 95-109.

Mathews, A. & Tucker, C. (2014). Government Surveillance and Internet Search Behavior. Retrieved from http://ssrn.com/abstract=2412564

Matzen Jr., R. N., & Sorensen, C. (2006). E-write as a means for placement into three composition courses: A pilot study. In P. F. Ericsson & R. H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 130-147). Logan, UT: Utah State University Press.

McAllister, K. S., & White, E. M. (2006). Interested Complicities. In P. F. Ericsson & R. H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 8-27). Logan, UT: Utah State University Press.

McCurry, D. (2010). Can machine scoring deal with broad and open writing tests as well as human readers? *Assessing Writing*; *15*(2), 118-129.

Meyer, J. & Land, R. (2003). Threshold concepts and troublesome knowledge: Linkages to ways of thinking and practising within the disciplines. *Occasional Report 4*, ETL Project. Edinburgh: University of Edinburgh.

Min-hsiu, T. (2012). The consistency between human raters and an automated essay scoring system in grading high school students' English writing. *Action In Teacher Education, 34*(4), 328-335.

Mogey, N., Paterson, J., Burk, J., & Purcell, M. (2010). Typing compared with handwriting for essay examinations at university: letting the students choose. *ALT-J, Research in Learning Technology*, *18*(1), 29-47.

Moxley, J. (2013). Big Data, Learning Analytics, and Social Assessment. *Journal of Writing Assessment*, 6(1). Retrieved from http://www.journalofwritingassessment.org/article.php?article=68.

National Council of Teachers of English. (2013). Position statement on machine scoring. Retrieved from http://www.ncte.org/positions/statements/machine_scoring

O'Neill, P., Adler-Kassner, L., Fleischer, C., & Hall, A. (2012). SYMPOSIUM: On the framework for success in postsecondary writing. *College English, 74*(6), 520-524.

Pagano, N., Bernhardt, S. A., Reynolds, D., Williams, M., & McCurrie, M. K. (2008). An inter-institutional model for writing assessment. *College Composition and Communication*, 60, 285-320.

Page, E. B. (1966). The Imminence of grading essays by computer. *Phi Delta Kappan*, *47*(5), 238-43.

Page, E. B. (1968). The use of the computer in analyzing student essays. *International Review of Education, 14*(22), 201-225.

Page, E. B. (2003). Project essay grade: PEG. In M. D. Shermis, & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 262-273). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Page, E. B., & Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan*, *76*(7), 561-65.

Penrose, R., & Gardner, M. (1999; 1989). *The emperor's new mind: Concerning computers, minds, and the laws of physics*. Oxford; New York: Oxford University Press.

Perelman, L. (2012a). Construct validity, length, score, and time in holistically graded writing assessments: The case against automated essay scoring (AES). In C. Bazerman, C. Dean, J. Early, K. Lunsford, S. Null, P. Rogers & A. Stansell (Eds.), *International advances in writing research: Cultures, places, measures* (pp. 121-132). Fort Collins, CO: The WAC Clearinghouse and Parlor Press.

Perelman, L. (2012b). Mass market writing assessments as bullshit. In N. Elliot & L. Perelman (Eds.), *Writing assessment in the 21st century: Essays in honor of Edward M. White* (pp. 425-438). New York, NY: Hampton Press.

Perelman, L. (2013). Critique (Ver. 3.4) of Mark D. Shermis & Ben Hamner, "Contrasting State-of-the-Art Automated Scoring of Essays: Analysis." Retrieved from http://graphics8.nytimes.com/packages/pdf/science/Critique_of_Shermis.pdf

Piccinini, G. (2007). Computationalism, the Church-Turing Thesis, and the Church-Turing Fallacy. *Synthese, 154*(1), 97-120. Retrieved from http://www.jstor.org/stable/27653444

Piccinini, G., & Bahar, S. (2012). Neural computation and the computational theory of cognition. *Cognitive Science*, *37*(3), 453-488. doi:10.1111/cogs.1

Quinlan, T., Higgins, D., & Wolff, S. (2009). *Evaluating the construct-coverage of the e-rater scoring engine*. (ETS Research Report No. RR-09-01). Princeton, NJ: Educational Testing Service.

Rawls, J., & Kelly, E. (2001). *Justice as fairness: A restatement*. Cambridge, MA: Harvard University Press.

Roscoe, R. D., Varner, L., Cai, Z., Weston, J. L., Crossley, S. A., & McNamara, D. S. (2011). Internal usability testing of automated essay feedback in an intelligent writing tutor. *Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference.*

Saygin, A.P., Cikekli, I., & Akman, V. (2000). Turing Test: 50 years later. Retrieved from http://crl.ucsd.edu/~saygin/papers/MMTT.pdf

Scheutz, M. (Ed.). (2002). *Computationalism: New directions*. Cambridge, MA: The MIT Press.

Schwegler, R. (1991). The politics of reading student papers. In R. Bullock, J. Trimbur, & C. Schuster (Eds.), *The Politics of Writing Instruction: Postsecondary*. Portsmouth, NH: Boynton/Cook Publishers.

Schunn, C. (2013). Learning from peer review: Current projects. Retrieved from http://www.lrdc.pitt.edu/schunn/research/peers.html#current

Scott, J. (1996). Postmodern gravity reconstructed, slyly. *The New York Times.* Retrieved from http://www.nytimes.com/1996/05/18/nyregion/postmodern-gravity-deconstructed-slyly.html

Searle, J. R. Minds, brains, and programs. *Behavioral and Brain Sciences 3*(3), 417-457.

Scott, T., & Brannon, L. (2013). Democracy, struggle, and the praxis of assessment. *College Composition and Communication*, *65*(2), 273-298.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, *27* (3), 379–423.

Shaughnessy, M. P. (1979; 1977). *Errors and expectations: A guide for the teacher of basic writing*. New York: Oxford University Press.

Shaw, E. J., & Kobrin, J. L. (2012). *The SAT essay and college performance: understanding what essay scores add to HSGPA and SAT*. Research Report 2012-9. The College Board.

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation*. New York: Routledge.

Shermis, M. D., Burstein, J. C., & Bliss, L. (2004). The impact of automated essay scoring on high stakes writing assessments. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Shermis, M. D., Burstein, J. C. & Bursky, S. (2013). Introduction to automated essay evaluation. In M. D. Shermis and J. C. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 1-15). New York, NY: Routledge.

Shermis, M. D., & Hamner, B. (2012). Contrasting state-of-the-art automated scoring of essays: analysis. Retrieved from http://www.scoreright.org/NCME_2012_Paper3_29_12.pdf

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9(4). Retrieved March 12, 2015 from http://PAREonline.net/getvn.asp?v=9&n=4.

Swain, S., & Le Mahieu, P. (2012). Assessment in a culture of inquiry: the story of the National Writing Project's Analytic Writing Continuum. In N. Elliot, & L. Perelman (Eds.), *Writing Assessment in the 21st Century* (pp. 45-68).

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics*. Boston: Pearson/Allyn & Bacon.

Thagard, P. (2005). *Mind: introduction to cognitive science.* (second ed.). Cambridge, M.A.: The MIT Press.

The Citation Project. (n.d.). What is the Citation Project? Retrieved from http://site.citationproject.net/

Turing, A. (1936). On Computable Numbers, with an Application to the Einscheidungsproblem. *Proc. London Math. Soc., s2-42* (1), 230-265.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind 49,* 433-460.

U.S. Department of Education, Office of the Deputy Secretary, Implementation and Support Unit. (2012). *Race to the top assessment: Smarter Balanced assessment report.* (Year One Report). Washington, DC: U.S. Department of Education.

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research, 10*(2), 157-180. doi:10.1191/1362168806lr190oa

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263-287.

Weigle, S. C. (2013). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing, 18*(1), 85-99. doi:10.1016/j.asw.2012.10.006

White, E. M. (2012). Afterword. In N. Elliot, & L. Perelman (Eds.), *Writing assessment in the 21st century: Essays in honor of Edward M. White* (pp. 495-503). New York, NY: Hampton Press.

Whithaus, C. (2013). In M. D. Shermis and J. C. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 1-15). New York, NY: Routledge.

Wilson, M. (2006). Apologies to Sandra Cisneros: How ETS' computer grading program misses the mark. *Rethinking Schools* 20.

Winters, L. (1980). *The effects of differing response criteria on the assessment of writing competence.* (CSE Report No. 131). Los Angeles, CA: Center for the Study of Evaluation, UCLA Graduate School of Education.

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.