THE ANDROID ENGLISH TEACHER: WRITING EDUCATION IN THE AGE OF AUTOMATION

by

Daniel Ernst

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of English

West Lafayette, Indiana

August 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Irwin Weiser, Chair
Department of English

Dr. Bradley Dilger
Department of English

Dr. April Ginther
Department of English

Dr. Harry Denny
Department of English

Approved by:
Dr. Manushag Powell

To teachers.

ACKNOWLEDGMENTS

I would first like to acknowledge the unwavering love and support of my parents, Marty and Beth, without which I wouldn't have finished. I am also deeply grateful for the guidance and companionship of so many faculty, staff, colleagues, and friends here at Purdue, especially but not limited to: Dr. Irwin Weiser, Dr. Bradley Dilger, Dr. April Ginther, and Dr. Harry Denny; the wonderful Heavilon Hall staff who helped me navigate the graduate school's bureaucratic hoops; my cohort members (and Mitch and Carrie), friends from across the department and from around town, and my guys in Kentucky; I am also indebted to the scholars, teachers, and statistics graduate students outside the English department, particularly Aining Wang, who carefully worked with me on this and other projects. Finally, I must express my gratitude for being able to live the life of the mind these past five years on such a great campus. Purdue has been a welcoming home.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

In an era of widespread automation—from grocery store self-checkout machines to self-driving cars—it is not outrageous to wonder: can teachers be automated? And more specifically, can automated computer teachers instruct students how to write? Automated computer programs have long been used in summative writing evaluation efforts, such as scoring standardized essay exams, ranking placement essays, or facilitating programmatic outcomes assessments. However, new claims about automated writing evaluation's (AWE) formative educational potential mark a significant shift. My project questions the effectiveness of using AWE technology for formative educational efforts such as improving and teaching writing. Taken seriously, these efforts portend a future embrace of semi, or even fully, automated writing classes, an unprecedented development in writing pedagogy.

Supported by a summer-long grant from the Purdue Research Foundation, I conducted a small-*n* quasi-experiment to test claims by online college tutoring site Chegg.com that its EasyBib Plus AWE tool can improve both writing and writers. The experiment involved four college English instructors reading pairs of essays comprising one AWE-treated and untreated version per pair. Using a comparative judgment model, a rubric-free method of writing assessment based on Thurstone's law, raters read and designated one of each pair "better." Across four raters and 160 essays, I found that AWE-treated essays were designated better only 30% of the time (95% confidence interval: 20-40%), a statistically significant difference from the null hypothesis of 50%. The results suggest that Chegg's EasyBib Plus tool offers no discernible improvement to student writing, and potentially even worsens it.

Finally, I analyze Chegg's recent partnership with the Purdue Writing Lab and Online Writing Lab (OWL). The Purdue-Chegg partnership offers a useful test case for anticipating the effects of higher education's embrace of automated educational technology going forward. Drawing on the history of writing assessment and the results of the experiment, I argue against using AWE for formative writing instruction. In an era of growing automation, I maintain that a human-centered pedagogy remains one of the most durable, important, effective, and transformative ingredients of a quality education.

CHAPTER 1: INTRODUCTION

"Education is not preparation for life; education is life itself."

—John Dewey

1.1 School as a Social Institution

A popular trope in science fiction is the blurring of the line separating human from machine. Typically, the trope reflects a fear that machines--broadly referred to as Artificial Intelligence (AI)--will become indistinguishable from humans as they become more human-like, and that constructing robots in our own likeness amounts to a fatal mistake. Underdiscussed, however, is the notion that the successive refinement of artificially intelligent machines cuts both ways, as historian and political scientist Adam Garfinkle suggests: "In science fiction, the typical worry is that machines will become human-like; the more pressing problem now is that, through the thinning out of our interactions, humans are becoming machine-like."

In an essay titled "The Erosion of Deep Literacy," Garfinkle ruminates on the impact of "pervasive IT-revolution devices" on our society and culture. He discusses a range of well-trod topics, both positive and negative, such as technology-induced short attention spans, the democratization of information made possible by the internet, and the communicative consequences of being always connected to the online world. Ultimately, he resolves on a note of caution:

> The more time we spend with machines and the more dependent on them we become, the dumber we tend to get since machines cannot determine their own purposes — at least until the lines cross between ever smarter AI-infused machines and ever less cognitively adept humans. More troubling are the moral issues that could potentially arise: mainly ceding to machines programmed by others the right to make moral choices that ought to be ours.

In this analysis, I see parallels to challenges facing education today. The "machines" of education include the deluge of apps and interactive virtual modules that simulate classroom lessons with increasing sophistication. As educational technology continues to be refined with ever-more "human-like" capabilities, educators, administrators, and policy makers are faced with serious questions about how much of it to incorporate into classrooms, courses, and curricula. Although technology enhances education in many ways, its overall impact on pedagogy remains unclear.

In American education, there is a long history of valuing the social and civic aspects of schools dating back to the theories of John Locke. John Dewey, the renowned twentieth century public intellectual and champion of progressive education, famously argued for schools as fundamentally "social institutions." The field of rhetoric and composition reflects that view. In her 1980 essay "Tacit Tradition," Janet Emig explains that Dewey is "everywhere in our work," connecting the inherent social and human character of composition to Dewey's theories of a social, civic, and progressive educational model (Fishman, 1993). However, the idea of school as a social institution is under threat by the inherent distance and flattening of human interactions that come with technology as a medium through which teaching is conducted.

Virtual and online courses currently have the most potential to contribute to this threat. Protopsaltis and Baum (2019) estimate that today nearly one-third of college students take courses online with no in-person component, and half of those students are enrolled in exclusively online programs. Recent disruptions to in-person classrooms, as well as over twenty years of online course offerings, provide plenty of data on the merits of online education. At best, online courses offer the advantage of asynchronicity and increased access and can sometimes achieve "no statistical difference" in learning outcomes than in-person alternatives

(Swan, 2003). At worst, online courses actively contribute to increasing gaps in educational success and disproportionately impact low-income and underrepresented student groups (Protopsaltis and Baum, 2019). Crucially, the biggest objectors to virtual education are often the students themselves, writes Peter Herman (2020), declaring in a recent Inside Higher Ed op-ed that "Online Learning is not the Future."

Extreme versions of virtual classrooms, those oversaturated by technology, reduce the humanity of education to a point where teachers and students become robotic; students interact with algorithms and modules more than with their teachers and peers, while teachers are demoted from educators to mere moderators. Most importantly, as Garfinkle notes in the passage quoted above, overreliance on educational technology and virtual classrooms could end up ceding to tech developers "moral choices" about curricular content, pedagogical approaches, and educational theories that ought to be locally determined by individual educators.

1.1.1 Covid-19 and the Turn to Online Education

It must be noted that the early 2020 emergence of the novel coronavirus, Covid-19, has inflected the conversation about education and technology with even greater importance. The Covid-19 health crisis has forced schools and teachers across the country to engage with students remotely through video applications and online instruction modules in order to comply with health experts' advice of social distancing. This is in addition to an already-growing popularity of virtual education. As this pandemic continues to alter in-person gatherings for the foreseeable future, creating a greater need for technological solutions, education stakeholders face serious dilemmas regarding the relative risks of traditional in-person classroom models and the limitations of virtual pedagogy. While this dissertation concerns the automation of education generally and the automation of writing instruction specifically, those issues cannot be totally

divorced from the concept of online education more broadly, as the latter functions as a through line to the former.

1.2 Automated Writing Evaluation

As questions involving the appropriate role of educational technology and virtual education mount, the teaching and assessment of writing occupies a unique position. College writing educators, administrators, and scholars have notable interests in this conversation because writing education is a fundamental tenet of the academy, which makes it particularly valuable. It is valuable because writing education enjoys a certain public and administrative mandate—everyone expects a college education to involve a significant amount of writing. Technological advancements in writing assessment have complicated this dynamic, however, with online services and educational technology companies competing to offer the newest, most cutting-edge writing education software, the latest of which is automated writing evaluation (AWE).

AWE refers to the use of computer programs to analyze and process human-produced writing. AWE is a catch-all category encompassing a range of technologies. Simple grammar checkers that come standard in word processors can be considered AWE, as can sophisticated Natural Language Processing (NLP) tools that conduct sentiment analysis. Within this spectrum, there exists a relatively new cottage industry of AWE tools that claim specifically to improve writing. These tools analyze written text using complex statistical models and algorithms and provide instantaneous feedback to users. Sometimes the feedback is as simple as prescriptive yes/no word-choice changes, while at other times it is more pedagogical, prompting the writer with questions about intended meaning and writing style. Crucially, the embrace of AWE as an

educational tool marks a significant break from traditional human-based writing pedagogy, and it is this break and the potential implications that are the subject of this dissertation.

1.3 The Challenge of Writing Assessment[1]

The academic field of writing assessment concerns the measurement and teaching of writing ability. The historical trajectory of the field can be understood as an attempt to cultivate ever more valid and reliable practices; that is, to produce forms of assessment that evaluate what they actually claim to, and that do so consistently. From multiple choice tests to constructed response essay exams to semester portfolios (Yancey, 1999), assessment models historically oscillate, trying to validly capture and reliably measure the ambiguous construct of writing ability (Huot, 1996). The introduction of computer technology has complicated this endeavor. Computers can do things humans cannot, such as process tens of thousands of words and sentences almost instantaneously. The potential applications of computational processing power to writing assessment are therefore staggering and seductive—but also massively disruptive.

The emergence of AWE writing tools designed to improve writing for the purpose of teaching, rather than simply evaluating it along a scale, marks an important disruption in both the educational technology industry and scholarly inquiry. The field previously focused on a related but distinct concept in Automated Essay Scoring (AES), which concerns a computer program's ability to *score* writing according to an established scale and which is typically used in standardized test settings. AES can be considered an "evaluative" technology, but its evaluation is limited to ascribing a numerical score to an essay. AES programs therefore make fewer

[1] "Assessment" is a broad designation referring to the study of measuring academic ability, writing and otherwise; writing evaluation is somewhat narrower, referring to the assessment of writing using either summative numerical scales or formative analytic feedback.

pedagogical claims, and much scholarship about the assessment capabilities of AES concerns simply the correlation between AES and human essay *scores*—an issue of the AES program's reliability. More recently, however, writing assessment scholars have become interested in what is known as the "consequential validity" of essay scores generally—the consequences of score use and interpretation—which has prompted deeper questions about whether AWE should be limited to scoring (AES) or if it has other valid assessment capabilities (Shermis and Burstein, 2013).

Prompted by questions of the consequential validity of essay scores, writing assessment scholars have turned back from an interest in what they call "summative evaluation" toward "formative evaluation" (Condon, 2013; Graham et al., 2015). Scholars have wrestled with these concepts since the 1960s and 70s (Scriven, 1966; Bloom, Hastings, and Madaus, 1971), but Horvath (1984) provides a succinct overview for those interested in these ideas from a rhetoric and composition perspective. Broadly defined, summative evaluation "treats a text as a finished product and the student's writing ability as at least momentarily fixed," which makes its purpose to judge and rank (p. 137). Conversely, formative evaluation "treats a text as part of an ongoing process of skills acquisition and improvement, recognizing that what is being responded to is not a fixed but a developing entity" (p. 137). In short, grading versus teaching.

Although these terms are pitted against one another, each informs the other and they are not easily disentangled, especially in the arena of writing. The ongoing learning of writing inevitably leads to improved and more highly ranked fixed examples of that writing ability. Where exactly summative evaluation ends and formative evaluation begins, and vice versa, is a vexing question for assessment scholars, one that colors the analysis of this dissertation and which will be more thoroughly discussed in chapter two.

Historically, however, summative evaluation has dominated. Evidence of its historical dominance in the assessment field is the ubiquity of holistic scoring in the teaching of writing, in both exam and classroom contexts. In the narrower context of AES specifically, computer programs have successfully obtained high correlation values with human raters using holistic scales (Dikli, 2006). While some consider this correlation proof of the value of these programs, others think the correlation between human and machine ratings indicates a hollowness to the act of ascribing holistic numerical scores to writing that has long been undertheorized (Huot, 1990). This realization perhaps more than any other has prompted a renewed interest in more analytic rather than holistic approaches to writing assessment, consisting of formative feedback and evaluation metrics other than numerical scores. Whether this return to formative evaluation for the purpose of teaching, such as analytic feedback, can be successfully replicated and automated by machines is the significant concern of this dissertation.

1.5 Statement of the Problem

The research gap this dissertation is intended to fill concerns the effectiveness of automated writing evaluation technology to formatively evaluate, and therefore teach and improve writing, and the potential relationships between the corporate purveyors of such technologies and American colleges and universities. In addition, this project situates the effectiveness of Chegg's EasyBib Plus AWE tool and the case of Chegg's partnership with the Purdue Online Writing Lab (OWL) within the larger history of writing assessment. I argue that (semi) automated writing pedagogy represents a misguided but logical next-step given technological advancements in AWE and the widespread availability of virtual learning alternatives to in-person education.

Use of AWE computer programs for writing assessment has traditionally been restricted to summative evaluation efforts such as scoring standardized essay exams, ranking placement essays, or facilitating programmatic outcomes evaluations. However, new claims about AWE's formative educational value mark a significant shift, particularly considering the automation of other sectors of the economy. The potential alignment of the automation of education with other automated sectors of the economy motivates this research.

The questions I intend to answer in this dissertation are of two kinds: those which I am able to investigate empirically and those which require an informed theorizing. The primary research questions are as follows, with the chapter(s) in which they are answered noted in parentheses:

1. What is the history of writing assessment in the fields of rhetoric and composition and educational research? How does AWE's pivot towards formative evaluation capabilities reflect and/or disrupt writing assessment's historical trajectory? (Chapter 2)
2. What are the potential limitations and applications of AWE technology related to formative writing evaluation? (Chapters 2 and 3)
3. Do formative AWE tools, such as Chegg's EasyBib Plus, improve the quality of college student writing? If so, how effectively? (Chapter 3)
4. What does the partnership between Chegg and Purdue University's OWL reflect about public-private partnerships in a corporatized American higher education system? (Chapter 4)
5. What aspects of writing pedagogy can be automated, if any at all? How does the reduction in human interaction--in both virtual and automated contexts--affect the teaching of writing? (Chapters 3 and 4)

6. What political factors contribute to debates surrounding educational technology and education automation? (Chapter 5)
7. What does the increasing automation of the American economy mean for the future of education? What stands to change for students, parents, teachers, and policy makers alike? (Chapter 5)

Technological advances have long been championed as "silver bullet" solutions to a variety of social problems, including to those facing education. Although the technology in question constantly changes, the narrative remains the same. From the television to the internet, the personal computer to the iPad, technology has promised to fill in the cracks in our education system through appeals to personalization, customization, and innovation. AWE and related attempts to automate writing instruction are no different. As courses and classrooms are increasingly mediated by technology amid public health pandemics and virtual learning opportunities, we should question whether the marriage between education and tech is borne out of technology's actual ability to improve learning outcomes or is instead the result of other interests.

1.6 Methodology

Historically, scholars in rhetoric and composition have employed a variety of seemingly disparate research methods. As Haswell (2005) notes, the study of writing and its teaching at times requires methodologies that are replicable, aggregable, and data driven (RAD), which generally implies quantitative research. Quantitative methodologies involve the gathering, analysis, and interpretation of numerical data and represent the dominant approach to social and behavioral research. These methodologies are associated with a positivist paradigm (Teddlie and Tashakkori, 2009). Stemming from critiques of the unquestioned and "received wisdom" of

positivist epistemology (p. 6), however, writing studies in more recent years has sought alternative methodologies. Qualitative research methodologies, which involve the gathering, analysis, and interpretation of narrative information, are often more aligned with constructivist epistemological paradigms that are increasingly popular in the social sciences today (Teddlie and Tashakkori, 2009), and include methods such as interviews, ethnographies, and discourse analyses (Schultz, 2006).

Many scholars have observed tensions between the epistemological paradigms undergirding quantitative and qualitative methodologies (MacNealy, 1999; Teddlie and Tashakkori, 2009). The positivist paradigm associated with the former assumes reality can be reduced to objective facts, represented numerically, while the constructivist paradigm associated with the latter believes reality is instead constructed and fluid and can only be described through narrative discourse (Alkove and McCarty, 1992). To complicate matters further, questions also persist about the purposes of what MacNealy (1999) deems "library-based," or traditional literary research, versus empirical research (p. 6-7), the relative merit and reputations of both, as well as the role of theory in such research (MacNealy, 1999).

It is my view that the concepts of writing, pedagogy, and education I aim to study here are fundamentally empirical phenomena. That is, writing competence is a cognitive trait and the ability to write is a learned skill, each observable and measurable. This conception of writing therefore requires an empirical investigation more closely related to positivist paradigms and quantitative methods employed in the social sciences, but this project requires some additional qualitative methods as well.

Traditionally, empirical research (especially quantitative) has been the purview of social science disciplines such as education, communication, sociology, anthropology, and political

science. The social sciences grew out of Enlightenment principles and represent an attempt to apply the scientific method to social behaviors and phenomena. The humanities, meanwhile, typically engage in aesthetic--rather than scientific--analysis (MacNealy, 1999). Yet, the scientific method is more embedded in humanistic research than we might realize. As MacNealy observes,

> the humanities and the sciences are not so disparate as many have been led to believe. Although some have argued that numbers simply aren't important or appropriate in the humanities, if a literary critic were to argue for a new interpretation of why Hamlet sent Ophelia to the nunnery and gave only one line from the play as evidence, the critic would be laughed at by responsible literary scholars. Thus, numbers are important to literary scholars (p. 5).

This dissertation, while conceived in a humanities graduate program, exists at the intersection of rhetoric and composition and education research and has been informed by coursework taken in Purdue's departments of English, Educational Psychology, and Statistics. I therefore attempt to meld traditional social science methods, such as experimental research and statistical analysis, with additional qualitative, humanities-style methods such as interviews and artifact analysis in a mixed-methods approach.

Famously designated a "dappled discipline" (Lauer, 1984), rhetoric and composition has always existed in a liminal space between the social sciences and the humanities. While the "rhetoric" side of the field tends to produce more library-based and qualitative research, the "composition" side often engages in empirical research using quantitative methods. As with much research concerning education and pedagogy, the purpose of this dissertation's research is to "demonstrate" (using the terminology of Reis and Judd described below) a correlational relationship, specifically that between writing quality and mechanisms of formative writing evaluation.

Reis and Judd (2014) describe three broad categories of empirical research: demonstration, causation, and explanation. The three categories can be considered each successively more detailed, with demonstration describing correlations of potential interest, causal identifying variables that can be manipulated to result in certain effects, and explanation determining why or how relationships between variables occur. Because the line between formative and summative writing evaluation is not clear cut, establishing a correlational association between changes in writing quality and use of AWE technology could indicate potential applications and limitations of AWE technology for use as formative, pedagogical tools. That is, writing that has been demonstrated to be improved (or worsened) by AWE technology suggests potential applications and limitations of such technology as a teaching supplement (or replacement).

Educational research frequently seeks to determine the most effective pedagogical interventions. That is, which kinds of teaching techniques, classroom configurations, homework assignments, and so on, result in the best academic achievement outcomes. Methods of measuring academic achievement, or assessment, represent a related area of contention in the pursuit to determine the most effective pedagogical interventions. This dissertation concerns both questions of pedagogical interventions related to the teaching of writing, as well as approaches to the educational assessment of writing ability. In particular, I attempt to demonstrate whether AWE technology is associated with changes in writing quality as perceived by college English instructors, and then determine what the association between changes in the quality of writing means for formative writing evaluation designed to teach writing competence and improve writing ability. While I ultimately do not attempt to fully explain *why* the change I observe occurs, I do use the results to inform a speculative discussion.

1.6.1 Mixed Empirical Methods

Teddlie and Tashakkori (2009) explain that a "third research community" outside of strictly quantitative and strictly qualitative methodologies exists, which is known as "mixed methods" (p. 7). A more recent development in scholarship, mixed methodologists "advocate the use of whatever methodological tools are required to answer the research questions under study" (Teddlie and Tashakkori, 2009, p. 7). In a mixed methods approach, categories like quantitative and qualitative or library-based and empirical are not pitted against one another, but instead seen as more or less appropriate for different aspects of research projects. In designing this project, I determined the questions raised by this dissertation can only be answered by a variety of research methods, including both quantitative and qualitative. For this reason, I employed a mixed-methods approach.

The central questions are empirical: are AWE programs effective at improving writing? If so, how effective? What are the potential limitations and applications of AWE programs as they attempt to formatively, rather than summatively, assess writing? In addition to these questions, which are designed to be answered quantitatively and experimentally, I also raise questions about the future role AWE technology could play on campuses and investigate the history of AWE's development. These questions are answered through informal interviews and artifact analysis, as well as historical research into the background of AWE technology and writing assessment more broadly. Altogether, this dissertation attempts to gather narrative, historical, experimental, and quantitative data about the effectiveness of AWE tools to improve writing in order to better speculate on its potential future role in higher education. The following three sections will briefly describe the methods through which such data will be gathered.

1.6.2 Experimental Research

One of this project's central questions considers the ability of AWE technology to improve writing. At base, this is an empirical question—how does AWE technology (independent variable) affect college student writing quality (dependent variable)? And relatedly, what is the association between writing quality and writing ability? Experimental research is defined as "the study of the effect of the systematic manipulation of one variable(s) on another variable" (Ary, Jacobs, Sorensen, and Walker, 2014, p. 28). Because experimental research aims to isolate and control variables as a way to examine empirical relationships, I determined an experiment is appropriate to help answer my primary empirical questions and I designed an experiment to test the capabilities of AWE.

To increase the internal validity of my experimental design, before conducting the experiment I elected to work with Purdue's department of statistics through its Statistical Consulting Service program. I was paired with a statistics graduate student, Aining "Anna" Wang, and her advisor, Dr. Arman Sabbaghi, who met with me and my advisor, Dr. Weiser, for an initial meeting about experimental design and analysis. I then continued to work with Anna until the experiment concluded. During our initial meeting, we discussed methods of analysis and experimental design and consulted Ary, Jacobs, Razavieh, & Sorensen's (2007) list of eleven threats to internal validity to ensure the experiment's design posed no major problems.

1.6.3 Interviews

To further contextualize the quantitative data yielded by the experiment, interview data was also collected. I adopted Blakeslee and Fleischer's (2007) "informal interview" model, which sits in between "spontaneous interviews" and "formal interviews." For scholars conducting ethnographies or observational studies, spontaneous interviews allow the researcher

flexibility to gather information without interfering too much in the setting. Conversely, formal interviews follow a rigid sequence of questions and aim for consistency across interview settings. I determined informal interviews were appropriate for my purposes. I prepared questions beforehand, but I was ultimately more interested having a conversation. The conversational structure provided flexibility in responses that offered more valuable information. As Blakeslee and Fleischer (2007) note, the combination of preparedness and flexibility makes informal interviews "probably the most common types of interviews qualitative researchers use" (p.133).

1.6.4 Purdue-Chegg Artifact Analysis

In addition to experimental and interview research, I also conducted an artifact analysis of the partnership between Purdue's Writing Lab and Chegg, an online college tutoring resource and the purveyor of the AWE technology used in the experiment. As Blakeslee and Fleischer (2007) write, an artifact is "essentially physical evidence that researchers examine to better understand the issues and people they are studying" (p. 117). The artifacts under examination in this dissertation include press releases from both Chegg and Purdue about their partnership, as well as press releases from Chegg about its acquisition of the WriteLab AWE technology. I also gathered information from Dr. Harry Denny, the Director of Purdue's Writing Lab, detailing the administrative politics of the Purdue-Chegg partnership.

This artifact analysis helps extend the study of the empirical effects of AWE technology on writing quality to a deeper analysis of how such technology, and the companies that own it, relates to institutions of higher education more broadly. Because part of this dissertation concerns the future roles of AWE technology in education and the experiment itself used Chegg's AWE tool, I deemed the Purdue-Chegg partnership an artifact appropriate for analysis, viewing it as a potential test case for similar public-private partnerships.

1.7 IRB

This research project was submitted for review by Purdue's Institutional Review Board (IRB) to ensure standard ethical research practices were followed. On 10 April 2019, the Purdue Social Science IRB emailed the principal investigator, Dr. Weiser, to confirm the project met the criteria for exemption under 45 CFR 46.101(b). A copy of the IRB confirmation letter can be found in Appendix A. Despite the exemption status, I opted to use a consent form when conducting the experiment with participants. The consent form can be found in Appendix B.

1.8 Chapter Summaries

This chapter provides an overview of my study and contextualizes each of the areas of interest. Automated writing evaluation (AWE), writing assessment, and online education are put into an initial conversation. I discuss the motivation for this research and detail my study design through a discussion of methodology. I also situate the Purdue-Chegg case within larger discussions about educational technology on college campuses.

Chapter 2 provides a history of writing assessment and a literature review of research involving AWE, chronicling its emergence within the writing and educational assessment movement specifically. Tracing the historical "waves of writing assessment" (Yancey 1999), I explain how AWE represents the next logical step as writing assessment pivots away from a focus on summative to formative evaluation. I argue the pivot to using machines for formative feedback signifies the beginning of automated writing education.

Chapter 3 details the Chegg essay experiment. In this chapter, I review more recent experimental research on the effectiveness of AWE as a formative educational tool. I then describe the experiment, following a traditional methods (participants, design, procedures, materials), results, and discussion (IMRaD) format. The discussion section interprets the

significance of the findings that the Chegg-treated essays were comparatively better only 30% of the time. I then discuss several possible explanations, including the inability of AWE programs to parse genre; the unsuitability of AWE for college-level writing; the "flattening" of the individual's writing voice in favor of the algorithm's; and the danger of "writing to the program," an analogue to "teaching to the test."

Chapter 4 details the recent partnership between Purdue and Chegg. I speculate on the reality and limitations of automated education for America's increasingly corporatized university. I also consider institutional political pressures at play in the Purdue-Chegg partnership and analyze strategies that preserve the principles and autonomy of the humanities and social sciences. Ultimately, I argue that human-centered education should be preserved amid technological advancements.

Chapter 5 concludes the dissertation and scrutinizes the possibility of automated education in light of the automation of other sectors of our economy, particularly what we would stand to lose from it. The age of automation, I argue, will bring with it peculiar rhetorical marriages between progressive-sounding politics and educational technology companies seeking to disrupt the traditional classroom model. Writing educators and scholars should be vigilant as these debates begin to circulate, and we should continue to gather data to support our claims.

Scholars in rhetoric and composition possess a unique expertise in the deeply human endeavors of writing, pedagogy, and higher education. Our work is necessary as the landscape on campus continues to change in step with advances in technology. Assessment in particular occupies a powerful position within education, since it determines what counts as right or wrong, achievement or failure. As educators—and thus assessors—we should scrutinize efforts to automate this essential component of pedagogy. These technologies therefore have the potential

to literally encode and reproduce linguistic, and thus cultural, bias into their operation, and ultimately degrade the rhetorical sensitivity offered by a human teacher. Research about AWE can help reaffirm the importance of human interaction to teaching and contribute to improving the overall quality of higher education.

CHAPTER 2: THE PARALLEL HISTORIES OF WRITING ASSESSMENT AND AWE

"Commerce is our goal here at Tyrell. 'More human than human' is our motto."

—Eldon Tyrell, *Blade Runner*

2.1 Introduction

In this chapter, I describe the history of the field of writing assessment, paying close attention to the emergence of automated writing evaluation (AWE) technology in the 1960s and its subsequent impact on assessment scholarship and writing pedagogy. Traditionally, the field's history has been chronicled as a series of successive "waves" during which disparate models and theories of assessment come to dominate. The primary tension as each wave succeeds the other is one between the different assessment models' reliability and validity, which in common parlance means the ability of the assessment instrument a) to consistently evaluate writing and b) the extent to which the assessment instrument actually evaluates the construct it purports to. I explain how attempts to refine AWE technology have mirrored these waves in parallel, with AWE developers aiming to validate their programs and achieve reliably high correlation coefficients with human raters on assessment tasks using holistic and summative evaluation scales.

I then argue an alternative lens through which to chronicle the history of writing assessment is the slow paradigmatic pivot from summative to formative writing evaluation. For much of writing assessment's history, scholars and practitioners have concerned themselves with summative evaluation—the valid and reliable scoring of writing, usually according to exam criteria and holistic numerical essay scales. The relatively recent advent of the portfolio model and questions about the consequential validity of essay scores, however, have prompted a

renewed interest in formative evaluation—assessment for the purposes of teaching. If history is any lesson, AWE technology will, and in many ways already does, attempt to replicate and automate formative writing evaluation techniques, once again positioning itself in parallel with the field. Through the provision of analytic feedback and the algorithmic modeling of specific writing traits, AWE's embrace of formative evaluation brings us finally to the current moment in which companies like Chegg offer AWE tools that claim to improve and teach students writing.

2.2 The Writing Construct

To understand the contours of the field of writing assessment, we must first understand what writing is in an academic context—how it is defined in terms of competence and what factors constitute that competence. Like other academic traits, writing competence is a "construct"; it has no physical properties that we could see or touch, and thus measure with an instrument such as a ruler. Instead it has been inferred—constructed—by observing variance in the quality of writing, variance in quality of both product and process, from classroom assignments to novels, essays, newspaper articles, letters, and so on.

Many educators struggle to formally define writing competence and instead take a cue from Supreme Court Justice Potter Stewart's famous explanation of obscenity, claiming good writing is something they "know when they see." Education researchers, and rhetoric and composition scholars in particular, however, are interested in a rigorous definition of writing competence *beyond* simply "knowing it when they see it," in order for teachers to better teach it and assessment instruments to better measure it. Others too, like Artificial Intelligence (AI) or Natural Language Processing (NLP) researchers, benefit from a rigorous definition of writing competence so they can program machines that better evaluate, produce, and process writing and (natural) language and writing.

For the purposes of this project, I am interested in how the writing construct is defined by education researchers in the fields of rhetoric and composition and educational psychology. Once we have a better understanding of how researchers in these fields define the writing construct, we can then explore how it is assessed, and finally determine how appropriately this knowledge is employed by AWE technology.

2.2.1 Definitions from Rhetoric and Composition Studies Position Statements

One place to find an authoritative definition of the writing construct is disciplinary documents. The major professional organizations of the rhetoric and composition discipline--the Council of Writing Program Administrators (CWPA), the National Council of Teachers of English (NCTE), and the Conference on College Composition and Communication (CCCC)--each have published relatively recent position or outcomes statements broadly detailing their definitions of the "writing construct." Taken together, these statements provide a snapshot of contemporary theories of writing, as well as its assessment and teaching.

The CWPA's 2014 "Outcomes Statement for First Year Composition" describes the "writing knowledge, practices, and attitudes that undergraduate students develop in first-year composition." Careful to term them outcomes and not standards, which the CWPA believes should be locally determined by individual writing programs and schools, the CWPA's definition of the writing construct combines three broad areas: 1) rhetorical knowledge; 2) critical thinking, reading, and composing; and 3) the processes by which we write. The first represents a subject knowledge commonly associated with writing, while the latter two detail how that subject knowledge is applied in different contexts and manifests through the use of various cognitive faculties and textual modalities. Note also that the CWPA devotes more space to defining the process of how students write more than the actual writing they produce.

The NCTE's 2016 "Professional Knowledge for the Teaching of Writing" offers a more general position statement than the CWPA's FYC-focused outcomes. The NCTE outlines several observations informed by decades of research on writing pedagogy. These observations cover a range of topics and include: writing is embedded in complex social relationships and their appropriate languages; composing occurs in different modalities and technologies; everyone has the capacity to write, writing can be taught, and teachers can help students become better writers; writing is a process; writing is a tool for thinking; writing and reading are related; and assessment of writing involves complex, informed, human judgment. The NCTE's observations overlap with the CWPA outcomes in significant areas. Both note that writing is related to thinking and that writing is fundamentally a "social" phenomenon. The NCTE does not define a subject knowledge area, like rhetoric, essential to writing, but it similarly acknowledges that writing is a complex process. Like the CWPA, the NCTE emphasizes writing as a learned action to be applied in various modes of academic inquiry.

The final position statement, CCCC's 2015 "Principles for the Postsecondary Teaching of Writing," in many ways combines the previous two. The most thorough of the three, the CCCC statement presents a "distillation of principles for sound instruction in postsecondary writing," which "extend from empirical research in the fields of English Language Arts and Composition and Rhetoric" as well as from "existing statements developed by the field's major organizations." Like the CWPA, the CCCC statement acknowledges the subject area of rhetoric as fundamental to the teaching of writing, and it also details writing's relationship to critical thinking, modalities, and writing technologies. Like the NCTE, CCCC notes that writing is a social act, is complex, and recognizes the importance of genres to the teaching and learning of writing.

In sum, these three position statements define the writing construct as a complex and iterative social phenomenon that combines critical thinking with a subject knowledge of rhetorical concepts, and which is ultimately assessed by the relative success of communication to various audiences in different genres and using different modalities. This definition is broad by design, allowing for individual teachers, programs, and schools to interpret it as they see fit for their own purposes.

2.2.2 Measuring Writing Competence

These position/outcomes statements define the writing construct primarily in programmatic terms, and therefore the resulting conception of writing competence does not lend itself to easy measurement, since programs can interpret them differently. How does one reliably measure a student's writing process, or a student's "capacity" to write? Defined as such, measuring writing competence risks becoming an assessment of the general cognitive ability of the writer, which only perhaps, to a greater or lesser degree, manifests in the writing itself in complicated, mediated, and approximated ways.

Scholars of classical test theory have long noted the difficulty in indirectly measuring abstract mental constructs (Crock and Algina, 1986), and writing assessment is no different. Measuring writing competence separate from general intelligence is deeply challenging, to be sure, but many educational psychologists and writing assessment specialists remain interested in the question of writing competence as an independent cognitive trait to be measured empirically. Crucial, then, to the successful measurement of writing is careful attention to the writing assessment instruments used. In other words, empirical investigations of writing competence as a unique cognitive trait depend significantly on the types of instruments used to test or measure it,

which is why studying the assessment of writing is so important to understanding writing more generally.

Due to the complex relationship between writing ability and the instruments that measure it, no one "true" test of writing exists. While it is possible to find different tests of writing whose scores correlate with each other, some scholars debate the extent to which the writing construct itself even exists independent of any specific assessment instrument used to elicit it (Ruth and Murphy, 1988). A writing assessment scenario typically consists of multiple components (the text produced, reading materials, the prompt), with both implicit and explicit features. These multiple components interact with multiple other factors, such as the instrument, the test taker, and the rater, and the assessment of the resulting product usually requires some amount of expert, yet still subjective, judgment to be made (Deane, 2013). All of these factors invariably shape our understanding of writing as well as the observed scores of writing competence an instrument yields. Ruth and Murphy (1988) provide a useful diagram displaying the multitude of mediating factors to consider in typical writing assessment scenario (Figure 1).

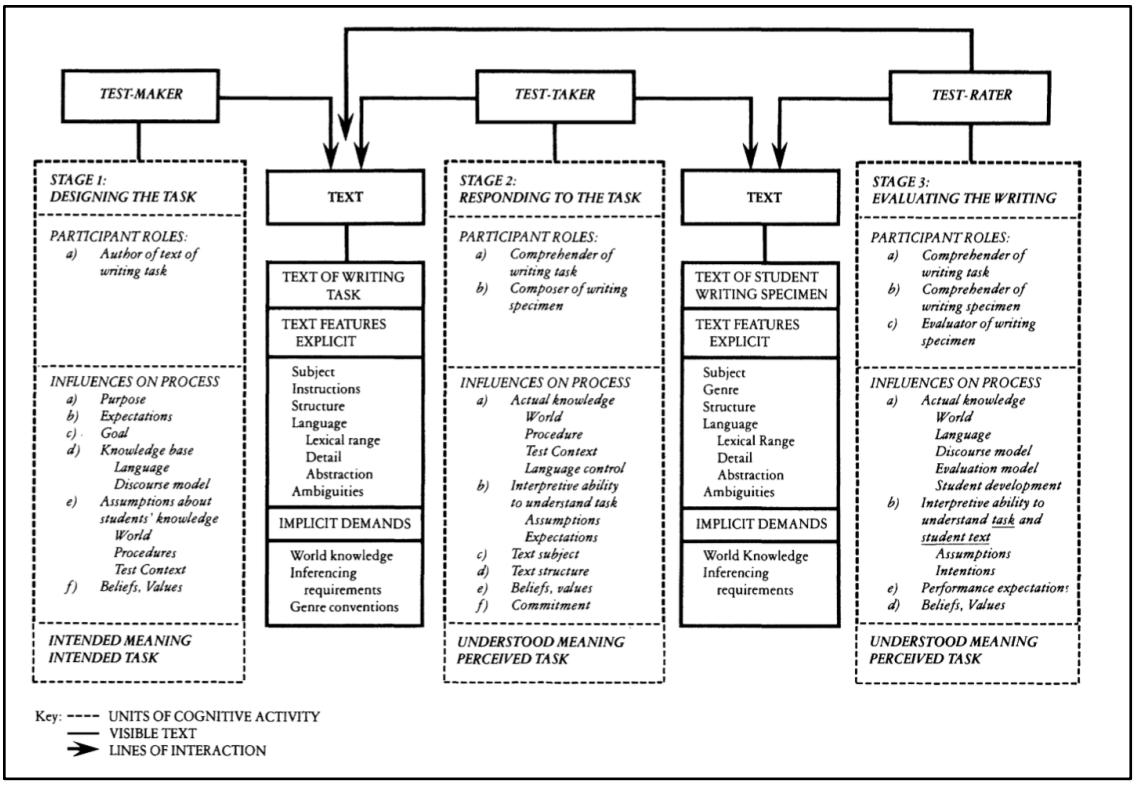


Figure 1. Factors Mediating Writing Assessment Scenarios

One way scholars in both educational psychology and rhetoric and composition have tried to simplify the complexity of writing assessment is by focusing on the most easily measurable features of the writing process, namely the text produced. But such a focus requires sacrificing attention to the more abstract elements of writing competence, those components outlined in the position statements, and instead emphasizes the more rote, mechanical aspects of writing. For example, an instrument that primarily assesses sentence level construction necessarily means that the writing *process* broadly construed is de-emphasized; it also requires a reduced emphasis on the social aspect of writing.

Many test instruments are criticized for simulating unrealistic writing situations. Proponents of authentic assessment (Wiggins, 2009) argue that standardized essay tests, for example, typically do not reflect common real-world writing scenarios, ultimately rendering

them counterproductive. Nevertheless, in order to facilitate some kind of consistent evaluation, assessment instruments of writing competence generally require some amount of standardization, and so they often attempt to assess observable features in a written text, such as its structure and the conventionality of language used. Diederich (1961) and Spandel's (1984, 2005) "six traits" of writing offer a useful summary of the traits of writing most instruments tend to emphasize: the successful employment of organization, ideas, voice, word choice, sentence fluency, and conventions. While broad, these traits can be located in the text produced by an assessment instrument and are thus more measurable than traits of writing competence located in the writer or process of writing. While it is true that mastering these traits does not necessarily make one a competent writer, they ostensibly serve as proxies for writing competence.

Perhaps it is the case that the assessment of writing competence is ultimately the correlative assessment of some other, deeper cognitive trait such as critical thinking. Or perhaps writing competence, like constructs such as *g*, or general intelligence, is a purely statistical phenomenon or a trait that is simply impossible to isolate and measure independent of other cognitive capabilities. The history of writing assessment and an overview of its theories can help us understand how scholars have approached the question of writing competence and its measurement, and it will also help us understand the potential for automated writing education in the future.

2.3 The History of Writing Assessment

2.3.1 Yancey's Three Waves

The assessment of writing and debates about how best to do it have been central concerns of rhetoric and composition since the field's inception. Although now taken up in ways unique to

the rhetoric and composition discipline, "writing assessment" was initially borrowed from the field of education and the social sciences more broadly, where it was studied alongside educational measurement and other cognitive psychology concepts. As such, early writing assessment efforts were largely based on the idea of objectively defining and measuring the writing construct through the use of multiple choice subject tests. As rhetoric and composition expanded and came into its own identity in the latter half of the twentieth century, it deepened its approach to writing assessment, and questions about valid and reliable assessment began to surface, such that we now acknowledge three successive periods in writing assessment's history. Kathleen Yancey (1999) has famously described these periods as overlapping "waves," spanning from the 1950s to the present, each articulating a particular theory of and approach to writing assessment. These waves provide a general map for navigating the history of writing assessment.

Yancey calls the first wave of writing assessment (1950-1970) the testing wave. This wave came out of the social sciences and is defined by the belief in the possibility of objective measurement of the writing construct through the use of multiple-choice tests of grammar and usage that purport to assess writing competence. During this wave, multiple choice writing tests were commonly used for classroom assessment, placement, and admissions, and some are still used today (see the ACT and SAT "verbal" sections). According to this theory, test items that assess the taker's ability to solve a variety of problems in subject areas relevant to writing such as punctuation, style, grammar, vocabulary, and similar, represent a measurable proxy metric for the ability to write.

Although such tests are highly reliable and easy to assess, they are not valid, critics argue, since they *indirectly* measure writing through proxy subject knowledges and fail to prompt students to produce writing (Yancey, 1999; 2012). Drawing on test theory nomenclature

(Crocker and Algina, 1986), multiple choice tests of writing subject areas lack *construct validity*, which concerns whether a test actually measures what it purports to. Questions about construct validity were the main motivation for scholars to re-think writing assessment beginning in the seventies, ushering in writing assessment's second wave.

Yancey's second wave of writing assessment is defined by the "holistically scored essay." Critics of the first wave's reliance on multiple choice tests of indirect writing knowledge asserted that to validly assess the construct of writing competence, we should instead directly assess actual writing (Yancey, 1999). The second wave therefore embraced constructed response items—essay tests—which proponents claim solves the problem of construct validity, since such instruments rely on direct assessment of self-produced writing (Huot, 1990). The essay test came to define the second wave, stretching from the l970s until the 1990s, and served as the standard form of writing assessment for classrooms, placement, and admissions.

However, essay tests came with their own pitfalls. While their use mitigated concerns about the construct validity of multiple-choice tests of writing, concerns about the validity of using a single essay exam arose in addition to assessment practitioners struggling to obtain reliable, or consistent, essay test scores. Whereas a multiple-choice test has only one correct answer per item, and thus a fixed score, one essay assessed by two readers can result in two different scores. Which score is correct? The unreliability of scoring essay tests precipitated writing assessment's third, and current, wave.

Yancey's third wave of writing assessment is defined by the writing portfolio. Use of writing *portfolios*, which combine multiple essays, essay tests, or similar constructed response compositions, is thought to address the concerns about essay exams' validity and reliability. While still utilizing direct writing assessment, the writing portfolio boosts both validity and

reliability by simply increasing the number of data points (essays); as the logic goes: if a single essay test provides insufficient construct validity, then assessing two (or more) essays is more valid than assessing one. Moreover, if two raters assign different scores to the same essay, then scoring multiple essays would theoretically reduce that unintended variance, resulting in a more accurate (or reliable) assessment of the student's true writing competence.

Portfolio assessments are common today, particularly in the assignment sequence of many writing-intensive classes. Some placement and admissions exams, such as the GRE and the TOEFL, assess at least two brief essays or multiple writing prompts, resulting in a "combined" analytical writing score similar to that of a portfolio. Indeed, scholars increasingly view portfolios as more accurate assessment instruments than standardized tests or other tools used for programmatic assessment (Sackstein, 2019). While the use of portfolios mitigates concerns about the unreliability of assessing a single essay, portfolios usually rely on holistic scales similar to those used in second-wave single essay tests and are thus subject to the same criticisms of holistic scales.

2.3.2 Automated Writing Evaluation: The Beginnings to Today

Although Yancey's history of the cascading waves of writing assessment's central theories is instructive, it fails to note a second-wave development crucial to the history and future of writing assessment. In the late 1960s, scholars began experimenting with then-new computer technology to evaluate writing, marking the beginning of automated writing evaluation (AWE). Broadly defined, AWE[2] refers to the use of computer programs to evaluate human-produced

[2] Sometimes referred to as Automated Essay Scoring (AES), Automated Essay Evaluation (AEE), or simply "machine scoring."

writing for a variety of purposes. Project Essay Grade (PEG), widely considered the first AWE program, was developed in 1968 by Ellis Batten Page to score student essays, for example.

But AWE efforts now range from simple tools such as grammar checkers and essay scorers to more sophisticated programs that perform functions like latent semantic analysis (LSA) and sentiment analysis. Unrestrained to the rhetoric and composition discipline and closely tied to advances in computing power, AWE has come to encompass a sprawling, cross-disciplinary effort to enlist the processing power of computers in all manner of evaluating human language. Between its second wave beginnings and Yancey's 1999 retrospective, AWE was largely dormant due to computational limitations, but beginning in the late 90s AWE has undergone a renaissance and continues to be of significant academic interest today.

There is a reason AWE has existed outside, if parallel to, the mainstream history of writing assessment. Since its inception, teachers and scholars alike, particularly those in the liberal arts, have routinely criticized AWE technology and its uses on the grounds that machines simply cannot replicate the efforts of a skilled educator in writing assessment (Attali, 2013; Haswell, 2006; Herrington and Moran, 2012; Condon, 2013; Deane, 2013; Perelman, 2014, among many others). AWE's most visible criticism surrounds its use in standardized testing. In 2013, a group of academics launched HumanReaders.Org, a petition to push back against the specific use of machine scoring of student essays written for high-stakes test instruments, with such notable figures as Noam Chomsky endorsing the organization and *The New York Times* reporting on its efforts. Later that year, the National Council of Teachers of English (NCTE) released an official position statement against the use of AWE, stating that, despite obvious savings in cost and labor, "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."

Despite the onslaught of criticism, the advent of AWE technology in the late 1960s and its continued refinement to the present represents a logical step--perhaps even moreso than the emergence of portfolio assessment--in the history of writing assessment. In contrast to Yancey, Brian Huot (1996) offers an alternative lens through which to view the history of writing assessment, wherein AWE's emergence is better accounted for. Instead of three waves, Huot argues the history of writing assessment is best understood as an oscillation between prioritizing the twin concepts of reliability and validity, with the first wave emphasizing the former, the second wave the latter, and the third wave both. Using this lens, we see that AWE emerges during a crisis in writing assessment reliability, which computers are uniquely suited to address. And until recently, AWE technology has largely concerned itself with whether computers can score essays as reliably as human raters (Kolowich 2012).

Today, however, new concerns surround not the ability of machines to score essays, but the value of essay scores themselves, irrespective of whether they come from humans or machines. In other words, critics now question the *consequential validity* of essay tests, which refers to the meaning and use of essay scores (Huot, 1990; Williamson, 2012; Shermis and Burstein, 2013). While in Yancey's history, questions of validity concerned the assessment construct--whether the instrument is actually measuring writing competence as it purports to--questions of consequential validity concern the very nature of assessment instruments and summative scores generally. Per Huot's oscillating history, a crisis in essay scores' consequential validity represents yet another vacillation in the history of writing assessment from reliability back to validity. In order to understand how these current debates about consequential validity

involve AWE specifically, we must first explore holistic writing evaluation, as well as the difference between summative and formative educational assessment.

2.3.3 Holisticism

According to its earliest pioneer, assessment scholar Ed White, holisticism is an attempt to evaluate "wholes rather than parts" and is the idea of assigning a single score to a piece of writing (White, 1984, p. 400). Holisticism rejects the piecemeal framework of analytic evaluation and multiple-choice tests in favor of evaluating pieces of writing as whole units, and in this way views a piece of writing as greater than the sum of its parts. Ironically, although White claims holisticism rejects the reductionism that plagues other methods of writing assessment, many critics have argued holistic scoring is itself deeply reductive (Charney, 1984; Vaughn, 1991; Elliot, 2005). By squeezing all the moving parts of an essay into a single score, holistic scoring inevitably prioritizes only those elements of writing that can be most reliably and consistently measured, such as grammatical correctness, usage errors, and basic features of organization (Haswell, 2006).

Despite this criticism, holisticism nevertheless helped solve the problem of unreliability that threatened the second wave of direct writing assessment. By norming readers to rate to specific scripts, holistic assessment instruments can achieve high reliability coefficients while at the same time allowing for test items with strong construct validity—items that prompt actual writing for direct assessment. This was a major achievement for writing assessment, and by the late twentieth and early twenty-first century, the popularity of constructed response essay items and the attendant paradigm of holistically assessing them was ascendant.

This ascendance is perhaps best exemplified in the 2005 changes to the SAT. Before 2005, the SAT had undergone years of criticism that it failed to measure abilities predictive of

college success, and one of its biggest customers, the University of California system, was particularly critical of its lack of an essay test (Llanos, 2005). In response to the criticisms, the SAT eliminated its verbal analogies section and replaced it with an essay test component, changing its famous 1600 point scale to 2400. At the time, this was seen as a step forward in improving the test, a modification that might promote equality in an instrument that many view as inherently biased. But such optimism was likely inflated largely due to the strong face validity[3] of essay tests at the time, but only ten years later the SAT made the essay test optional and reverted back to the 1600-point scale, perhaps mirroring growing concerns elsewhere about the consequential validity of holistic essay scores.[4]

During the ascendance of holistically scored essay tests, AWE was similarly ascendant. This is because AWE technology can easily imitate the same simple judgments required of human raters using holistic scoring frameworks. Over the last couple of decades, researchers have refined AWE technology and produced a growing body of literature confirming AWE's reliability at such tasks, proving machines can rate as reliably as their human counterparts, and sometimes even more so (Dikli, 2006; Kolowich, 2012). By demonstrating that AWE scores correlate very highly with those of trained human raters, corporations and universities have justified the continued and widespread use of AWE in contexts such as standardized testing and course placement, making it common today (Dikli, 2006).

[3] Face validity is how valid a test appears in terms of its stated claims, or its "authenticity." Face validity is often discussed in terms of "task-based" or "real-world" assessment and learning.
[4] The UC system abandoned the SAT/ACT requirement altogether in 2020, further hinting that the current crisis has much to do with the consequential validity of tests.

2.3.4 Correlating Essay Scores

The issue of correlation—and reliability more generally--has been the primary focus of AWE developers for the last couple of decades. While many AWE programs continue to be refined and updated to this day, some have already been adopted by various corporations and government agencies to be used for both research and business purposes, as well as for statewide and even national educational assessment (Smith 2018). Their primary function, however, remains limited to the *scoring* of essays, particularly in large scale, high-stakes assessment situations.[5] Yet, because the challenge of scoring essays reliably has largely been met, current efforts to refine AWE technology today represent an attempt to move computers beyond their ability to score essays to a more granular evaluation of specific traits of writing, similar to how a teacher might assess writing. In effect, current and future progress in AWE development reflects a desire to move from rote, summative scoring to scrupulous, and human-like, formative writing evaluation designed for the purposes of education.

Most AWE programs operate using fairly simple statistical methods. ETS's e-rater, for instance, relies on linear regression for score prediction (Burstein et al., 2013). Linear regression is a statistical analysis technique in which the relationship among quantitative variables is examined, specifically how criterion (dependent) values vary by changes in predictor (independent) variables. For automated essay scoring, the summative, holistic score of an essay is thus vital, since techniques like regression require all inputs to be in the form of quantitative variables. Computers are primed first by being fed corpora of edited text and large data sets of essays prescored by humans, and regression analysis allows the programs to identify recurring

[5] Otherwise known as "standardized tests."

linguistic patterns and analyze basic features of the best and worst rated essays as determined by human readers.

Programs then use that information to predict how new essays might score comparatively (Larkey and Croft, 2003). For instance, a program might count the number of subordinate clauses in an essay and compare it to the number of subordinate clauses in essays highly rated by humans. The regression analysis first tells the machine how big of a role the number of subordinate clauses plays in explaining the variance in scores of the human-rated essays and then predicts its own a score accordingly. Once enough variables are combined, such as word length and type/token ratio, and multiple linear regression models with different variables weighted more or less heavily than others are used, these programs begin to score essays similarly to and as reliably as trained human raters (Larkey and Croft, 2003).

Linear regression is still used today, but techniques have evolved thanks to improvements in computing power as well as advances in methods of statistical modeling. Building on the idea of multiple linear regression, researchers now commonly use a technique known as Latent Semantic Analysis (LSA) that models essays as vectors, such as in the KAT engine program. As Williamson (2009) explains, LSA is a "dimensionality-reduction" method that transforms the content of an essay into a vector with two variables, length and direction. The program positions various essay-vectors in particular places in a multidimensional virtual space and then compares their locations and directions, based on the similarity of linguistic features, to those of human prescored essays. The direction of one essay is compared to that of another by the degree of the angle separating the vectors, and this value is weighted and averaged together with vector length and then combined with other linguistic measures to yield an essay score.

But all the correlational research and sophisticated statistical models obscure the more salient story. No matter how accurately machines can be programmed to rate writing like human counterparts, the most vocal critics argue that high correlations between human and machine *scores* do not mean that humans and machines *read* texts the same way. In fact, whether you can even call what machines do to calculate their scores "reading" is itself the main question, since some critics consider it mere "counting" (Herrington and Moran, 2012; Perelman, 2014). Amid these disputes, the primary critique remains that, despite a congruency of scores, machines can't "read," much less assess, writing in a way at all similar to an expert human reader. If we accept this critique, then human and machine assessments mean inherently different things, and this renders a congruency between human and machine scores useless, since a machine simply imitates or calculates, but never replicates, what human assessment actually yields.

This line of criticism also calls into question the very process of summatively scoring essays itself, even if performed by a human. In other words, it cuts both ways: if a computer can be programmed to score like a human, then the human can be said to score like a computer. The advent of reliable machine scoring, while convenient and efficient, reveals a deeper crisis of the consequential validity of holistic essay scoring more generally. But AWE developers have not let that deter them. As Shermis and Burstein (2003) noted more than fifteen years ago, "the direction of automated evaluation of student writing is beyond the automated prediction of an essay score" (p. xiv). So what, exactly, is this "direction"? Are AWE developers seeking to capitalize on a crisis in the consequential validity writing assessment models?

2.4 From Summative to Formative Evaluation

The new debate surrounding AWE's consequential validity represents the latest site of contention in the history of writing assessment (Kane, 2013). Much as essay tests emerged to

alleviate concerns with the construct validity of multiple-choice writing tests, and holisticism emerged to mitigate the unreliability of assessing written essays, AWE researchers are now attempting to refine AWE programs to address concerns of consequential validity. To do so, AWE is moving away from holistic scoring frameworks in which whole essays are assigned a single numerical score and attempting to model more specific, and natural, "traits" of writing. Additionally, researchers are incorporating Natural Language Processing (NLP) techniques into their AWE algorithms to better capture and model qualities of writing beyond the mere "countable" features (Shermis and Burstein, 2003, 2013). This pivot from reliable holistic scoring to natural trait modeling provides a final lens through which to view the history and future of writing assessment: as an evolution from prioritizing summative to formative writing evaluation.

Viewing the history of writing assessment as an evolution from prioritizing summative to formative evaluation better enables an understanding of AWE's role going forward. Critiques of the consequential validity of holistic scoring initially gave rise to portfolios, whose multiple component parts are thought to reduce the variance inherent in holistic assessment by examining larger, more varied and representative samples of student writing competence, which increases reliability and validity. But the embrace of portfolio assessment also indicates a higher order shift beyond questions of consequential validity and reliability. Portfolio assessment places a greater emphasis on the formative, rather than purely summative, evaluation of student writing competence.

Portfolio assessments generally include teacher feedback, as well as drafts of compositions, so that a greater focus on writing development and process are emphasized, as is a range and variety of writing traits, which is reflected in the portfolio model's inclusion of

multiple assignment genres. We can therefore view the portfolio paradigm as one in which formative evaluation is prioritized over summative evaluation—students compile a portfolio not just to be scored, or graded, but so they can see how their writing ability forms and evolves over the course of a semester.

2.4.1 Modeling Traits

As a new paradigm with greater emphasis on formative writing assessment emerges, AWE programs designed solely to score written essays have become outdated. AWE researchers and developers are therefore adapting to keep up. In a return to a more analytic—as opposed to holistic—model of the writing construct, many AWE programs now target the assessment of particular traits of writing. "Trait assessment" is less concerned with assigning summative numerical scores and more interested in providing formative evaluation, such as the provision of evaluative feedback, the identification of genre characteristics, and the analysis of rhetorical features and style.

Automated feedback that goes beyond summative numerical scores combines advanced statistical modeling with cutting-edge techniques in machine learning, natural language processing (NLP), and artificial intelligence. Some of these programs are designed with singular goals in mind, like the LIWC2015, which purports to model a writer's "affective profile" through textual analysis. Other programs and researchers are interested simply in improving machine modeling of human language for its own sake, waiting for applications of such models to reveal themselves down the road. Nonetheless, developers appear mostly unified in the attempt to design AWE programs that can do more than score essays.

Likely influenced by potentially widespread and highly remunerative NLP applications in AI, current AWE research concerns a range of complex questions: How can computers more

closely "read" text rather than "count" it? Can computers understand figurative language, such as metaphors, or make analogic inferences like human readers? These and other research questions, I argue, contain significant implications for writing educators and students alike, as traditional human-centered formative pedagogy of writing hinges on similar questions. The closer to human readers computers become, the greater the potential for changing the way writing is taught and learned.

Since essay scores alone provide little-to-no formative information for learning writers, AWE programs are increasingly providing feedback like a teacher might. Teacher feedback on student writing in the early stages often comes with no numerical score, or even comments on sentence level grammar issues, and instead focuses on what we might call a composition's "rhetorical" qualities: how well the piece meets genre requirements, the overall organization, the persuasiveness of the argument, paragraph development, and so on. How might an AWE program intervene in these areas?

Some programs, such as the Biber tagger, claim they can identify and analyze an essay's genre. As Crossley et al. (2015) explain, the tagger compares the incidence of certain textual features in an essay, from word level to clause level, to that of its nearest generic counterpart to formulate a generic match between essays. This allows the user to see how well they are mirroring generic features. Another program, the Coh-Metrix, is reported to assess overall clarity of writing by modeling "text difficulty, text structure, and cohesion" (McNamara et al., 2015, p. 39). The program uses metrics such as word-concreteness, word familiarity, clausal comparisons, measures of lexical diversity, and overall essay cohesion to assess that trait, offering the user feedback on clarity of thought like a teacher might in a conference with a student. These programs thus claim to provide formative evaluation capabilities at the level of a composition's

rhetorical features, such as genre and clarity. Critics might argue that these capabilities are genre-bound, and any feedback provided remains severely limited compared to a human evaluator, but nonetheless the very attempt to do so suggests this is the direction future AWE is headed.

While parsing genre remains one of AWE's biggest challenges, some researchers are attempting to tackle this problem by designing programs that identify and analyze common rhetorical structures of writing that transcend genre. Rhetorical Structure Theory (RST) informs recent attempts to identify and sort essays based on, for example, their introductions and conclusions, thesis statements, evidence and support, etc. (Burstein and Marcu, 2003), which are essay features that tend to transcend genre and are commonly found in most pieces of writing.

As for providing a premium beyond the subjective evaluations of human evaluators, a computer providing this kind of analysis has vast implications for writing instruction in particular, since a machine that can reliably identify such structures "would permit one to provide *automatic* feedback about the presence or absence of discourse elements, the quality of each of those elements, and the strength of the connections between discourse elements in an essay," as opposed to the gradual accumulation of writing knowledge students experience in a traditional classroom (Shermis and Burstein, 2003, p. 219, emphasis mine). An *automated* component of this kind of formative feedback lends an efficiency to writing pedagogy that human teachers cannot match in processing ability; they can only hope to provide superior formative feedback than computers, which is where this debate has currently stalled.

A computer's ability to correctly identify an essay's conclusion or thesis statement, or determine its genre or "clarity of thought," is still not the same as an expert educator's careful evaluation of that same essay. Can a computer assess, for instance, how persuasive one argument is versus another? Or how successfully a writer makes use of metaphor, or irony, or other

nonliteral language tropes? Can a computer process the affective tone of a heartbreaking personal narrative?

Surprisingly, some answer yes to those questions, or at least think a "yes" is one day possible. Isaac Persing and Vincent Ng, researchers at the Human Language Technology Research Institute at the University of Texas at Dallas, have made recent attempts to model constructs they call "strength of argument" and "stance of writer," which were previously thought too complex to model (Persing and Ng, 2015, 2016). These models, which use a regression analysis of features of the text that supposedly capture the writer's argument, claims, and opinions, are designed ultimately to calculate a strength of argument measure to predict essay performance, and initial results are promising.

Similarly, researchers are also attempting to model and capture the use of nonliteral language like metaphors and analogies (Fass, 1991; Gedigan et al., 2006; Mason, 2004; Shutova, 2010; Heintz et al., 2013; Hovy et al. 2013; Huang 2013; Klebanov et al., 2014; Klebanov et al., 2015; Lamb, 2017). Programs designed to detect and interpret metaphors use an array of methods, such as those similar to the tagger program for metaphor identification and algorithms to interpret metaphorical meaning (Shutova, 2010). More importantly, these efforts represent the increasing combination of AWE methods with those of NLP, a natural coalition when considering AWE's attempts to move beyond essay score prediction to the assessment of writing traits—"natural" features of language rather than the purely synthetic numerical scores holisticism assigns to essays.

In sum, the current state of AWE is in transition. As computer programs have come to score essays just as reliably as human raters, the usefulness and validity of the scores themselves are now being questioned. Going forward, it appears AWE aims to identify and evaluate

particular writing traits, and to process them more deeply than simply assigning them a numerical score.

2.4.2 Future Applications of Formative AWE

Many aspects of trait-focused writing evaluation pose challenges to automated machine models, but perhaps the simplest way to summarize the challenge is the question of whether machines "read" or "count" text. Les Perelman (2014) has famously argued that computer writing assessment programs do not read, they count. "Counting" textual features is less of a problem if the end result is a numerical score, but for trait-focused writing evaluation, counting may fail to provide useful, or consequentially valid, feedback.

The comprehension of nonliteral language exemplifies one of the principal, and seemingly irreconcilable, differences between a human understanding of written text and that of a machine. Human writing rarely, if ever, uses solely literal language for extended periods of time; we frequently draw on metaphors and analogies to help convey our points. But "counting" figurative linguistic constructions proves difficult for computers and becomes especially problematic when a metaphor or analogy is crucial to the successful interpretation of written text.

In the nomenclature of AWE developers (Filiz, 2018), this problem concerns Natural Language Understanding (NLU). Closely related to Natural Language Processing (NLP) and Automatic Speech Recognition (ASR), NLU represents the narrowest, and most challenging, aspect of automated language evaluation. MacCartney (2014) shows how the three terms are related and how success in NLU could result in a program capable of much narrower evaluation of language, allowing for trait-based feedback on the rhetorical, "uncountable' dimensions of writing like the use of nonliteral language tropes (Figure 2).

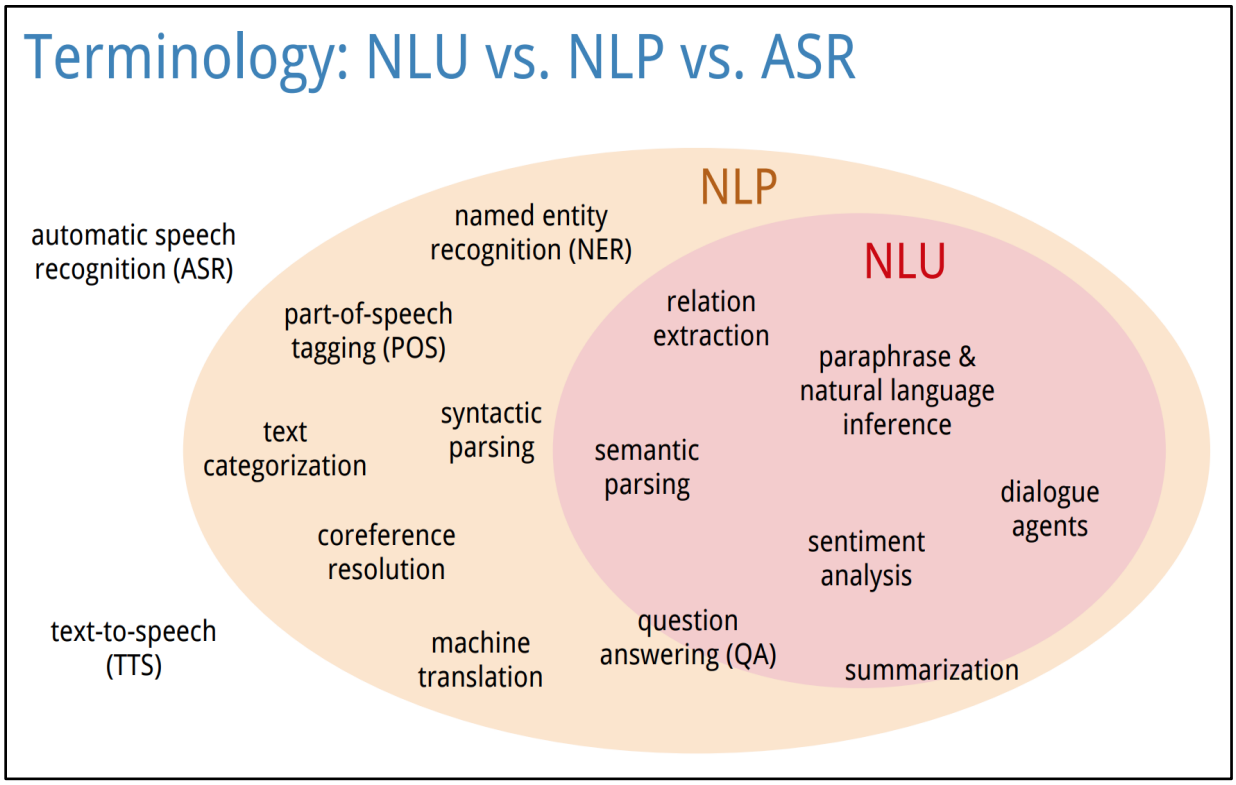


Figure 2. Distinctions Between NLU, NLP, and ASR

Whether the combination of NLP and AWE as well as refinements in statistical modeling could result in a machine program that truly "reads" or "understands" written text is perhaps a more philosophical than empirical question. Nonetheless, the practical applications and potential of machine programs—for educators and students alike—that sufficiently approximate the human ability to formatively evaluate, analyze, and assess natural language traits is staggering.

2.4.3 Supplement or Supplant?

Education scholars and educators have long debated the merits of Computer Assisted Instruction (CAI), since the emergence of the personal computer in the 1980s, and our pedagogy continues to be mediated by computers more and more every year. But due to computational limitations, computers have always been treated as supplemental aids. AI and AWE are prepared

to revolutionize CAI, especially with the widespread accessibility of such tools on the internet. While many scholars urge caution and point out that AWE should function solely as a *supplemental* tool to the primary pedagogical efforts of human teachers (Ware, 2011; Wang, 2015), AWE technology that is sophisticated enough, combined with budget cuts and the increasingly popularity of online education, could come to *supplant* rather than supplement. Indeed, some scholars are already pointing to research that shows computer-mediated instruction increasingly outperforming more traditional instruction methods as evidence that AWE has greater potential than is now being used (Chandrasegaran et al., 2005; Shermis et al., 2010).

Although many scholars continue to assert that AWE should not and will never fully replace teachers, findings in the research and changes to our increasingly computer-mediated classrooms suggest AWE could potentially play a bigger role in the teaching of writing (Williamson, 2004; Cotos, 2011). Depending on how refined these programs become, arguments could soon be made for AWE technology—in certain contexts—to fully supplant, rather than simply supplement, writing teachers or writing-intensive classes. As Williamson (2004) advocates: we should "study automated assessment in order to explicate the potential value for teaching and learning, as well as the potential harm" (p. 100). If AWE technologies become refined enough that they are perceived to "read" as well as humans, it's not unimaginable to think that writing teachers could be replaced by AWE programs and algorithms, especially considering the potential labor and resource costs saved. Williamson (2004) continues:

> We have to expect that the future will hold some developments that can help us and some that can be hurtful. Some developments will be faddish, oversold by developers and producers of the technology, whereas others will enter our toolbox with the potential to help students learn if used properly (p. 100).

Looking back at the trajectory of the role computers have played in education, as well as the last 40 years of AWE development, it might be wise to ask: is the android English teacher the writing teacher of tomorrow?

2.5 From Grading to Teaching

Thus far, I have argued that the pivot of AWE technology away from essay scoring toward the evaluation of specific writing traits represents a logical step in the history of writing assessment. The primary issue of contention among writing assessment scholars today is the consequential validity of essay scores, which afflicts both machine scoring and human holistic scoring alike. The crisis of the consequential validity of holistic assessment has prompted a renewed interest in more analytical, trait-based assessment, giving rise to the portfolio and similar process-focused assessment instruments.

This literature review suggests in addition to Yancey's three waves, and Huot's pendulum metaphor of the field oscillating between a focus on validity and reliability, a new lens through which to view the history and future of writing assessment is the slow march from prioritizing summative (essay scoring) to formative evaluation (assessment of writing traits). This lens better accounts for the history and future of AWE, which unfolds parallel to, but distinct from, Yancey's waves and falls somewhat outside of Huot's analysis of the oscillation between validity and reliability.

I want to suggest one final way of viewing the summative-formative evaluation evolution, and that is as a transition from *grading* to *teaching*, a variation on summative and formative evaluation. Efforts in AWE development mirror the renewed paradigmatic focus on formative evaluation of writing by programming machines to teach writing, rather than just "grade"—or summatively rate—it. This is, finally, what brings us to the possibility of the

android English teacher, wherein writing instruction itself is automated with students writing to algorithms.

The jump from automated scoring to automated instruction is the result of significant conceptual shifts in recent years. As Condon (2013) discusses, simple assessments like a machine scoring an essay, though "correct" in the sense their scores correlate strongly with those of human raters, are ultimately simply "not very helpful" (p. 101). He continues: "Whether scored by humans or machines, these systems of writing are subject to the fallacy of surrogation—the substitution of a statistical artifact—a number—in place of the need for complex information" (p. 101). Whether that number comes from a human or machine is irrelevant.

Part of the problem I have been gesturing towards is that scholars and educators often confuse summative and formative educational assessment for one another, particularly when it comes to writing pedagogy. Assigning a paper a grade or scoring a placement test essay—summative evaluations—do not help teach or "form" struggling writers into shape; summative evaluations simply rank-order students relative to one another or an established scale (A-F). Conversely, providing careful written feedback on a student essay—formative evaluation—does not necessarily provide a generalizable picture for the student of how their writing competence stacks up, but it does (presumably) help grow student achievement (Trumbull and Lash, 2013).

The formative/summative dynamic is played out in a number of related contexts. In K-12 education circles, the debate is framed as "growth versus proficiency" (Lachlan-Hache and Castro, 2015). In the world of educational testing, scholars distinguish "norm-referenced" from "criterion-referenced" tests (Crocker and Algina, 1986). Nonetheless, the essential dynamic remains: when we assess we can sort and rank student performance relative to each other and

group averages on scaled tests to get a sense of overall performance, or we can assess and foster individual student growth according local criteria like specific class, or programmatic, outcomes. Both are necessary, and indeed each helps the other, but in the context of writing assessment particularly, and especially when debating the merits of AWE, we must be clear about what kind of assessment the technology as it exists permits, and what it claims to be able to do in the future.

However, some scholars caution against being too reductive in their conception of potential uses for AWE, particularly scholars in Second Language Studies (Rich and Wang, 2010). Weigle (2013) reminds us to make important distinctions among the populations who might use AWE technology, such as second language learners both in their native home or abroad; distinctions between writing-to-learn and learning-to-write; and distinctions between instruction (formative) and assessment (summative). With these distinctions in mind, and noting that the validation of language tests now increasingly takes the form of ushering arguments (Kane, 2013), Weigle and others make the case that AWE could be beneficial for second language learners and beginner writers, especially given the potential for immediate feedback.

The benefits of AWE for second language learners, as well as struggling student writers, depend entirely on how we *use* the technology. Currently, we use AWE mostly as a summative evaluation tool, to sort students into performance bands and place them into courses, graduate programs, and so on, and given the logistics of processing hundreds of thousands of student applications and tests, it seems the cost-benefit analysis in this realm works in its favor.

However, now that computers are as reliable as humans at numerically rating and ranking essays, the question has naturally progressed to whether AWE can be used for formative evaluation purposes: to help teach, for example, second language learners or new college students how to write. What would that mean for current and future college teachers? How might

AWE technology interface with increasingly popular online education offerings? Would this mark the beginning of fully automated education, where even the teacher is a machine? Finally, and more importantly, can a computer really teach us how to write?

Today, the combination of computers and language is ubiquitous beyond the world of educational assessment, with Natural Language Processing (NLP) and Natural Language Understanding (NLU) at the forefront of breakthrough technologies like personal assistants and necessary for the successful future development of Artificial Intelligence (AI) programs. As computing technology has advanced, and the scoring component of writing assessment been mastered, interest in machines' formative evaluative capabilities and their pedagogical potential have become the new focus and future of writing assessment. No longer is assessment primarily concerned with the summative assignment of scores to writing, but instead in the formative analysis, understanding, processing, and teaching of writing and natural language, such that writers can be provided with feedback and instruction on how to improve their writing. This shift has enormous implications for students and teachers of writing and presents an important area of future research for scholars.

2.6 The Emergence of Web-Based AWE

This literature review has explored the progression in AWE from an emphasis on programs designed to summatively score writing to those designed to formatively evaluate it. This progression naturally prompts questions about the current and future use of programs designed as such. I believe AWE will be used more frequently in pedagogical capacities, eventually resulting in the semi-automated—and potentially fully-automated—teaching of writing. However, institutions of learning will be slow to embrace automated teaching technologies. Like the automation of other human-centered tasks, the public remains skeptical

and the idea of "android teachers" likely offends many. Moreover, whether these programs are effective at teaching writing remains an open question.

Questions about the effectiveness of AWE programs as teachers should be investigated empirically. If we look at the rise of free and cheaply available internet-based AWE tools in the last decade, we are provided a vast laboratory to test AWE programs before their institutional endorsement. Web-based AWE programs vary in scope; some, like Grammarly, amount to little more than a grammar-checker. Others, however, are more ambitious, seeking to cement themselves as the go-to student aid for writing instruction. As interest in "computer-generated feedback" grows, it is more important than ever to get out in front of the technology and collect empirical evidence on the abilities and limitations, affordances and conveniences of formative AWE technology so educational institutions can make informed decisions about its adoption.

One example of a popular web-based AWE tool is The Hemmingway App. As a tool, it is extremely easy to use and efficient. A user simply copies and pastes text into the module and it offers instantaneous feedback, along with calculating the reading level, wordcount, and readability rating (from poor to good). The automated feedback focuses on things like word choice, passive-voice construction, and density and clarity of sentences. The feedback it provides is fairly simple, counseling writers to use fewer adverbs and less passive voice, for example, but the feedback is largely diagnostic, offering very few prescriptive solutions. Students and teachers alike could conceivably benefit from such a tool, as it is so easy and efficient to use. But would the tool actually improve the writing and increase student learning? And, even if it does, is the improvement worth the risk of a tool like this replacing or at least significantly altering the feedback from a teacher?

Questions about the program's effectiveness or the pedagogical value added of AWE tools are increasingly topics of concern for scholars. Paige Ware (2011) examines how much computer-generated feedback helps improve student writers. Ware concluded that the answer depends on how writing is defined; that is, whether the writing is defined as a narrow set of mechanistic skills or a broader expression of critical and rhetorical thinking. Defined narrowly, Ware believes computer-generated feedback can benefit student writers, but crucially notes that the studies that showed improvement involved intensive and long-term use of the AWE programs. Additionally, she acknowledges that many teachers worry about the kind of writing AWE programs teach: mechanistic, formulaic, and divorced from real-world contexts. Ultimately, she cautions that questioning the effectiveness of computer-generated feedback may be the wrong question all together. Instead, she argues:

> Teachers would be advised to critically analyze the cost-benefit relationship of its use depending on the features of writing that are considered important by the particular institutional and instructional constraints in which they make pedagogical choices. Over the long term, effects on the more observable mechanistic and formulaic aspects of writing may be counterproductive if they lead teachers further away from writing purposefully for real audiences (771).

Ware urges restraint about the ability of these programs, and presumably most teachers, and probably the public too, would agree with that urge. But, many of the AWE tools emerging today do not display a similar level of restraint when it comes to claims about their ability.

Perhaps due to the competitive nature of the educational technology market, many AWE programs make grand claims about their tool's abilities. Chegg, arguably the most popular online-tutoring resource for college students today, is one of the latest to join the AWE race. After acquiring WriteLab, an AWE program developed by UC-Berkeley PhD student Matthew Ramirez, for $15 million, Chegg appears to be touting its ability to help improve college student writers. Chegg, and WriteLab before them, have lauded the program's sophistication,

emphasizing buzz terms like Artificial Intelligence and Natural Language Processing. In press releases about the acquisition and profiles about Ramirez, the tool is carefully differentiated from rudimentary grammar checkers by promoting it as something beyond "squiggly red and green lines" (Sternlicht, 2018). Instead, it is said to provide lessons in "substance and style" and not only "offers revisions and edits, but also poses questions and suggestions, encouraging students to learn by making decisions about what they want to say and how they want to say it" (Chegg, 2018). Whether Chegg's AWE tool can improve college student writing is an empirical question, and the next chapter details a quasi-experiment designed to help answer just that.

2.7 Conclusion

This chapter chronicled the history of the field of writing assessment. I detailed how, throughout this history, AWE technology has followed a parallel arc to that of writing assessment generally, with AWE developers attempting to validate and make more reliable technology that replicates human writing evaluation. Crucially, the field of writing assessment has recently pivoted from interests in summative to formative writing evaluation, prompted by questions of holistic scores' consequential validity and the popularity of the writing portfolio assessment model.

Using the history of writing assessment as a guide, I argue that developers of AWE technology will mirror this pivot, refining AWE technology to replicate and automate formative writing evaluation techniques such as analytic feedback and writing trait modeling and identification. This pivot has potentially significant implications for the teaching of writing, since formative evaluation is more closely associated with pedagogy than summative evaluation. In other words, this chapter demonstrates that the automation of writing education is not far off, and

we will be wise to begin gathering data on its limitations and applications sooner than later, as the next chapter attempts to do.

CHAPTER 3: THE ANDROID ENGLISH TEACHER AND THE CHEGG ESSAY EXPERIMENT

> "A man barely alive. Gentlemen, we can rebuild him. We have the technology. We have the capability to build the world's first bionic man. Steve Austin will be that man. Better than he was before. Better, stronger, faster."
>
> —*The Six Million Dollar Man*

3.1 Introduction

In this chapter, I conduct a small-*n* quasi-experiment to test the effectiveness of a popular online automated writing evaluation (AWE) tool (Chegg's EasyBib Plus) at improving college student writing. The experiment is designed to gather data about the formative evaluation capabilities of an automated computer program that claims to possess pedagogical faculties aimed at improving both student writing and learning. This experiment was conceived of and conducted within the context of discussions of writing assessment generally and AWE specifically as outlined in chapter two.

From Microsoft Word's grammar check to Educational Testing Service's (ETS) Criterion tool, AWE technology has an extensive history dating back to the 1960s beginning with Ellis Page's Project Essay Grade (PEG). AWE programs range in sophistication, from ranking a set of essays or assigning them scores to providing feedback on writing with revision suggestions that improve grammar, clarity, and style. Although AWE technology has improved significantly over the years, it has generally been limited to surgical uses—for specific summative evaluation purposes like admissions, placement, or programmatic outcomes evaluation.

Now, however, AWE developers make bolder claims, for instance that their programs are "instructional" or "educational" and can help students become better writers (Purdue University News, 2019; Educational Testing Service). Although seemingly innocuous, such claims evoke

"silver bullet" rhetoric and reach far beyond more modest previous assertions of a computer's ability to score an essay along a scale or rank a sample of essays according to machine-learned criteria. Instead, claims of instructional and educational value represent a sharp pivot from the realm of summative to formative evaluation, which marks a major shift that could bring significant implications for students and teachers alike, especially with the automation of other sectors of the economy looming on the horizon.

3.2 Literature Review

3.2.1 Pedagogical Interventions for Formative Writing Evaluation

As AWE attempts to offer formative writing instruction, it is imperative we compare its effectiveness to that of existing formative instructional methods. Formative writing instruction currently exists in many evaluative configurations. Most traditionally, teachers provide detailed feedback on student writing, instructing students using a series of drafts, feedback, and revisions. In secondary and higher education, students also commonly receive feedback from peers (peer review) as well as engage in self-assessment or revision. On college campuses, writing centers regularly offer one-to-one formative consultations to help students become "better writers" (North, 1984). While AWE is not yet as popular as any of the aforementioned formative evaluation mechanisms, it is increasingly common as a third-party, web-based resource for students to employ at their own discretion.

Scholars debate whether computer-assisted pedagogy, let alone fully automated writing evaluation, can play an effective role in improving writing and how it compares to existing alternatives (Shermis, Burstein, & Leacock, 2006). Graham, Hebert, and Harris (2015) compared writing evaluation mechanisms in a meta-analysis of the effectiveness of four different feedback

sources for student writers in grades 1-8: adults, peers, self, and computers. They found that all feedback sources resulted in improved writing (average weighted effect sizes: adults, .87; peers, .58; self, .62), but that computers resulted in the smallest positive effect of the group (.38). In a case study examining the use of the Writing Roadmap 2.0 AWE technology for formative feedback, Rich and Wang (2010) also found positive effects of AWE use on middle and high school writers in the US, with low-performing students in particular seeing the greatest improvement.

Although computer-assisted formative feedback often yields positive effects in experimental research like this, there are some concerns. First, the effect of computers tends to be significantly smaller than human-centered alternatives (Graham, Hebert, and Harris, 2015). Another issue is that most research involves K-12 education (Myers, 2003; Landauer et al., 2009; Mao et al. 2018), and whether a computer program can improve college-level writing remains under-researched.

The studies on the effect of general computer-assisted teaching, not just writing instruction, at the college-level that do exist have been somewhat inconclusive. In an early meta-analysis, Kulik, Kulik, and Cohen (1980) showed computer-assisted college instruction made small contributions to course achievement. In an experiment involving students in an online teacher education course, Riedel et al. (2006) found use of automated essay scoring (AES) programs improved the quality of student papers based on final scores by human raters; however, because the study involved students in an online course, no in-person pedagogy served as a control or comparison group.

While education researchers, particularly at the K-12 level, appear cautiously optimistic about the potential applications of AWE technology, AWE is largely criticized among rhetoric

and composition scholars who primarily study college writing and adult literacy. Their criticisms can be categorized into four primary areas of concern: the theories/constructs of writing undergirding AWE programs; inappropriate uses of AWE technology; the consequences of computerized writing assessment for classrooms, students, and teachers alike; and the impact of such technologies on underrepresented student populations (Elliot et al., 2013). To be sure, recent advances in AWE computational models do not mitigate such criticism. As AWE tools have been refined through the addition of Natural Language Processing (NLP) and Latent Semantic Analysis (LSA) techniques, scholars (Elliot, et al., 2013) remain critical because these advances do not allow the programs to "read" student writing like a human, but rather enable them to produce evaluative comments based on statistical modeling of measurable aspects of the text.

This latter concern—whether machines can really "read" text—represents the primary theoretical question motivating much of the criticism of the theories of writing undergirding AWE programs. Generally, psychometricians and computational linguists view language ability as a cognitive trait and writing as a measurable skill, whereas rhetoric and composition scholars tend to view language and writing as rhetorically complex and socially embedded behaviors. Strict adherence to the former view, argues Condon (2013), can result in teachers assigning the kind of writing that can only be evaluated by machines, such as short, timed essays of a specific genre, which fails to accurately reflect true student writing ability. Deane (2013) similarly argues that for AWE programs to succeed in the future, they must be programmed to account for writing's social and cultural dimensions. Otherwise, AWE systems will simply reinforce outdated and narrow notions of formal correctness that fail to account for modern, expanded definitions of the writing construct (Vojak, Kline, Cope, McCarthey, Kalantzis, 2011).

Scholars also criticize the uses of AWE programs. While most AWE programs are used, at least primarily, for scoring essays, critics argue that essay scores are only a small aspect of writing assessment and that numerical scores represent inherently reductive proxy measures for more nuanced writing abilities. Balfour (2013) showed that when used as a source of writing feedback for students enrolled in MOOC courses, AWE failed to improve writing compared to having students participate in peer review.

Nevertheless, some scholars (Cope, et al. 2011) argue that AWE has the potential to move beyond summative evaluation—scoring—to play a role in formative evaluation, or pedagogy. The central tension between AWE's role in summative versus formative evaluation centers around how much human teachers supplement, or are supplanted by, the technology. Deane, Williams, Weng, and Trapani (2013) see a role for AWE in large-scale assessment tasks of classroom activities, but they maintain that such tasks must be limited in scope to the measurable and technical aspects of writing. This position—that computers can only ever play a limited role in classroom instruction—appears to be the consensus among rhetoric and composition scholars.

In addition to concerns about the writing construct and the uses (or misuses) of AWE programs, rhetoric and composition scholars have also criticized the consequences of using automated writing evaluation. Consequences range in scope, from fundamental changes to classroom dynamics to alterations to curricula to the revamping of course placement and school admissions processes. All manner of literacy education—and some aspects of school administration and admission—stand to be affected by the proliferation of AWE technology. While some scholars urge writing program administrators to accept the inevitable and strategically adopt the use of some AWE programs (Klobucar, et al., 2012), others (Cheville,

2004) warn the increasing influence of automated writing educational technology could transform teachers into "data managers" and redefine writing instruction away from a transaction between writer and reader to a formula to follow by students.

Finally, some scholars express concerns over questions of diversity as they relate to AWE programs. While this issue has not yet been studied as thoroughly as the above issues, initial studies suggest AWE programs may have "disparate impacts based on gender, ethnicity, nationality, and native language, privileging or penalizing some cultural backgrounds and languages over others (Eliot, et al., 2013). However, some research regarding AWE use among English language learners and students with disabilities, specifically, shows potential benefits for these specific groups. For example, scholars have theorized that the instantaneous feedback capability of machines is well-suited for L2 students or English Language Learners (ELL), pending further validation (Ranalli et al., 2017).

In a meta-analysis of 37 studies comparing foreign language teaching supported by AWE technology versus pedagogy not supported by AWE technology, Grgurovich, Chapelle, and Shelley (2013) found a small but positive effect of using AWE technology in foreign language teaching. Other studies, however, have proved more inconclusive. Wang (2013, 2015) found that AWE programs helped L2 writers only in certain areas of formative evaluation like error analysis of usage and feedback on organization, and that students ultimately preferred a combination of human and machine feedback.

Students with disabilities (SWD) is a group scholars think could stand to benefit from AWE technology. Wilson (2017) compared growth in writing quality between SWD and typically-developing (TD) students when both groups used AWE programs. Results showed a positive association between AWE use and growth in writing quality for SWD, suggesting these

technologies may help close achievement gaps in the realm of disability. Overall, rhetoric and composition tends to view AWE programs, along with their use and consequences, with skepticism, urging educators to approach their use cautiously, but they remain interested in specific applications of the technology for certain groups of students.

3.2.2 The Case of Chegg

Most research about AWE tends to involve the proprietary programs of companies like ETS or Pearson, but as the technology has become more widespread, we are seeing greater access to AWE programs via online sources. For this reason, I have chosen to experiment with an online AWE program. Chegg.com, one of the most popular online tutoring hubs for college students, offers their own AWE program, the EasyBib Plus. Known for its online tutoring across all disciplines, as well as textbook rentals, Chegg has become a go-to resource for many college students, and state of the art writing assistance is one of the company's latest efforts. Chegg's EasyBib Plus is not a typical, freely available online AWE tool. As discussed in chapter two, Chegg acquired the AI-enhanced WriteLab AWE technology in 2018 and has since folded it into its writing pedagogy services. WriteLab was a sophisticated and respected program. At the time of its acquisition, "more than 500 U.S. classrooms utilize[d] WriteLab, with some teachers going so far as refusing to read a non-WriteLab-reviewed paper" (Sternlicht, 2018).

Besides its massive popularity, I also chose to analyze Chegg because of its recent partnership with Purdue University. Purdue's Online Writing Lab (OWL) is a world-famous resource for writers, students, and teachers. One of the earliest online writing labs, the Purdue OWL over the last 25 years has become an authority on all matters of writing, today generating approximately 40 million pageviews per academic year. In 2017, Chegg approached the writing lab at Purdue about a partnership wherein Chegg would license OWL content and monetize ad

revenue from the OWL. In addition, staff at the OWL would play an advisory role for Chegg. Such public-private partnerships are not uncommon, especially at large universities.

While this partnership seems common enough, and both parties stand to benefit, it portends a potentially bigger role for automated educational technology on college campuses in the future. For example, as Chegg's EasyBib Plus tool continues to be refined, how might the program affect the Purdue Writing Lab's in-person writing consultations or its online consultations? With this in mind, I conducted a mixed methods, quasi-experimental case study to test Chegg's claims and gauge the relative value of using AWE to improve and formatively evaluate student writing.

3.3 Methods

3.3.1 Overview

In summer 2019, I designed and conducted a small *n* experiment.[6] The experiment tested Chegg's claims that its EasyBib Plus program has the ability to improve student writing quality and learning outcomes by comparing writing treated by the EasyBib Plus to untreated drafts. Although only the EasyBib Plus program was tested, some findings may apply to the automation of writing feedback at the college level generally, independent of specific AWE tools. The goal of this experiment is to analyze an increasingly popular instructional supplement for student writers, one that Chegg, a massively popular college student tutoring site, has invested significant money in. Though limited in size, scope, and generalizability, the experiment opens pathways to broader questions about the limitations and applications of automated writing evaluation and identifies potential pitfalls for automated education more generally.

[6] The research project was IRB approved.

3.3.2 Participants

Four graduate student English instructors participated as raters. At the time of participation, all instructors had a minimum of three years of experience teaching college English; all had experience teaching their own sections of the English 106 course for which the essays used were written; all were PhD students in the English department at Purdue University; and all had taught, or were currently teaching, some form of argumentative writing in their classes.

3.3.3 Design

This study is a within-subjects (or repeated measures, meaning all participants are exposed to both treatment and non-treatment conditions) quasi-experiment. True experiments contain the most internal validity because they utilize a truly randomized sample of subjects, which best controls for confounding variables. However, Ary et al. (2014) observe that in much educational research, true experiments are impossible due to the ethical limitations of randomization of students: "neither the school system nor the parents would want a researcher to decide to which classrooms students were assigned" (p. 339). Therefore, quasi-experimental designs are often used in educational contexts, which attempt to randomize subjects as much as possible within given curricular constraints. Such is the case in this study; since I had no control over whether the student essays used were a truly random sample, I can only claim that the sample of essays is quasi-random.

The quasi-experimental design is fairly simple. Each of the four raters were randomly assigned 25 pairs of student essays from a sample of 85 pairs, 20 of which were unique to their sample and 5 of which were shared and evaluated by all four raters as a control mechanism. Each pair contained two versions of the same essay written by the same student, with one per pair

treated by the EasyBib Plus program. The raters were instructed to read each pair of essays sequentially and designate one of the pair "better"[7] in order to determine the frequency with which the essays evaluated by the EasyBib Plus are perceived as better than their untreated counterparts.

The "which is better" method of evaluation draws on Thurstone's law of comparative judgment. Thurstone's law is a model used in pairwise comparisons to measure differences in perceptions. It describes a technique commonly found in educational and psychometric research, as it attempts to isolate non-physical traits or attitudes of the mind: "the law is applicable…to qualitative judgments such as those of excellence of specimens in an educational scale" (Thurstone, 1994, p. 266). A comparative model is particularly appropriate for this experiment because it does not require the use of an assessment rubric; use of a detailed rubric risks measuring how well the rubric is applied to the evaluation of an essay rather than the determination of which essay is perceived as better.

At every stage of sample preparation, materials were randomized. The 85 essays were randomly selected from a population of 100 Purdue students across five different English 106 courses. Five essays were randomly selected as a control sample, to be evaluated by all four raters; the remaining 80 essays were randomly divided into four groups of 20 and randomly assigned to the four raters. When the raters were given the single stack of printed essays, the order of which essay (treated or untreated) appeared first was also randomized.

The experiment was designed for a point estimate analysis, which is the estimation of an unknown population parameter value. Point estimates are values between 0 and 1 that represent a

[7]The deliberately ambiguous term "better" was used in order to focus the experiment on the program's claims, which are similarly vague. In addition, by using a vague term across four different raters, the experiment better assesses the program's abilities rather than how well the raters would have applied specified criteria.

probability, in this case the probability a Chegg-treated essay is designated "better" than its untreated, paired counterpart. Since this experiment involves only one of many possible samples, a 95% confidence interval was calculated to produce a *range* of probabilities that the treated essays are perceived better by the raters, a more accurate estimation of the program's true success rate due to inherent sampling error. The range of point estimate values is tested against a null hypothesis value for statistical significance. In this case, the null hypothesis assumed a point estimate value of $\hat{p}$=.50, or 50%; in other words, the null hypothesis assumes that the probability of a treated or untreated essay being designated better is equal, like a coin flip. This would mean that the EasyBib Plus program provides no discernible improvement or detriment whatsoever to the writing quality of the essays. If the estimate does differ from $\hat{p}$=.50, we need to know if it was likely due to chance, so the point estimate would then undergo a paired sample t-test to see if it differed significantly from the $\hat{p}$=.50 null value, either positively or negatively.

3.3.4 Materials

The essays used were collected from my own first year composition (FYC) courses, five different sections of Purdue's English 106 (20 students per course, 100 essays total from which 85 were randomly selected) taught between Fall 2015 and Fall 2017. At the time the essays were written, the student writers varied by year, but most were first- or second-year students majoring in a variety of disciplines. The 85 selected essays thus represent a quasi-random and representative sample of (early) university students, since English 106 is a general education requirement and one of the most widely taken classes on campus.

Furthermore, all essays were taught using the same assignment prompt and rubric and were written at approximately the same point in each semester—as the second assignment in the sequence of major projects. This helped provide further control among the samples, since each

rater was rating papers that were written by students at roughly the same point in the semester, despite their coming from five different semesters. Papers written by students at the end of the semester versus the beginning might vary significantly in quality, introducing unwanted variance into the experimental results. The primary difference between essays from different sections in this sample, then, is only whether they were written in a Fall or Spring semester. All essays were completely de-identified and otherwise unmodified except to match typefaces, font sizes, and spacing. A numerical code in the top left corner was assigned to each essay, so only the researcher could identify the treated essays.

Chegg's AWE tool, EasyBib Plus, was used to treat the essays. EasyBib Plus offers instantaneous feedback on issues of style, grammar, writing clarity, and plagiarism (Chegg.com). Use of Chegg's EasyBib Plus program required me to manually enter each of the 85 essays into the program and accept grammar and style change suggestions at my own discretion.[8] All essay data was saved in an excel spreadsheet, and as the essays were treated, certain attributes were logged for later comparison to ensure each of the four samples contained essays of roughly equal quality and length. These attributes included average word counts of both treated and untreated essays, changes in word counts between the treated and untreated essays, number of edits made from the AWE revision suggestions, and ratios of changes in words per edit (see Table 1). These attributes helped determine if re-randomization was needed to ensure each sample was roughly equal to one another and the overall sample.

[8] This retro-treatment of essays is a limitation of the study but was done due to time constraints. A more rigorous study would have students use the program themselves. Even though I can only ever approximate how the program might be used, I attempted to use it consistently and not in a way that would disfavor the program.

Table 1. Essay Attributes for Each Rater Sample and Overall

Average	Words, Untreated	Words, Treated	Change in words	Edits	Word change per edit
Sample A	910.05	889.85	-20.20	29.90	.70
Sample B	926.95	908.80	-18.15	29.25	.64
Sample C	919.60	897.70	-21.90	29.25	.77
Sample D	904.50	887.10	-17.40	27.40	.66
Overall	915.28	895.86	-19.42	28.95	.69

During this process, I tried to accept the program's suggestions as often, but also as realistically, as possible. For instance, I tended not to accept any change that would too greatly alter or obscure the original meaning of the text. I tended not to accept changes that appeared to be blatant errors, like changing verb tenses from what naturally sounds correct. For the suggestions that required judgment calls rather than accepting simple yes/no revisions, I attempted to insert myself into the essay as little as possible to resolve the identified error; for example, the program frequently suggested resolving sentence fragments by "adding in missing information or combining the sentence with a nearby sentence," and I always tried initially for sentence combination. This element of the experiment is imperfect, as it is impossible to replicate how individual students might use such a program in their own way. However, I tried to engage with the program as honestly and consistently as possible.

The final material consideration for the experiment was which kinds of essays to use. Argumentative writing was decided on because it is a genre of writing commonly taught in FYC that often emphasizes issues of style, something many AWE programs claim to address. Editorials were chosen as a specific genre of argumentative writing because they are, again, commonly assigned in writing classes and relatively brief.

3.3.5 Procedures

The experiment was conducted on Purdue's campus over the course of four weeks in July and August 2019. After raters read and signed a consent form (see Appendix B), I provided some basic background information about the assignment for which the essays were written. Careful not to influence how the raters should evaluate the essays, I simply tried to give the raters an idea of what the assignment looked like in class. I explained the assignment was an editorial, and students were encouraged to write about timely topics with no right or wrong answer—to argue a position or offer an opinion on a topic and support it. I explained that students were assessed not only on the coherence of their arguments, but also on the style, clarity, and persuasiveness with which they wrote.

The procedures of the experiment were then explained (see Appendix C): In front of each rater was a single stack of 25 pairs of essays (50 total). Their job was to read each pair in the order presented, or to read the pair simultaneously, and designate one of each pair "better." They did not know that one of each pair had been treated by an AWE program, and they did not know that the order of each pair (that is, which essay of the pair appeared first) was randomized between the treated and untreated essays. Each pair was given a letter and numerical code during the de-identification process (for example, A29 and A34, B16 and B12), and after reading each pair, they entered the numerical portion of the code for whichever essay they designated better into a Qualtrics survey. They were allowed to take breaks as needed. Each of the four raters took between 2-3 hours to complete this portion of the experiment.

After each rating, I conducted a brief interview, lasting approximately 5-10 minutes. During these interviews, I tried to get a sense of the criteria each rater used/developed for designating one of the pair better, since I did not enforce any prior criteria. The purpose of these interviews was to learn how the raters interpreted the differences in the essays, and also to

determine if they approached the task with significantly different mindsets. These interviews, in addition to the shared sample of 5 essays that all raters read, served as a mechanism to determine if the results should be analyzed in total (n = 80), as if each rater were interchangeable, or independently (n = 20), as if each rater represented its own probability. In the interest of full disclosure, I have included the results of both analyses in the section to follow.

3.4 Results

3.4.1 The Essay Experiment

The experimental results can be analyzed two ways. In the first, the analysis considers the total sample overall (n = 80) by assuming each of the four raters are interchangeable—that they approached the rating task similarly enough that the majority of potential irrelevant variance is controlled for by experimental design and randomization. The advantage of this mode of analysis is the larger sample size, which provides greater confidence in the results. To gauge the viability of this analysis, all raters read the same five pairs of essays at the beginning of their sorting. This shared sample (n=5), as well as the post-experiment interview during which raters were asked questions to better understand their individual processes and assessment criteria were compared for noticeable differences. In the shared sample, no rater designated Chegg-treated essays better more than 40% of the time, suggesting a similarity in their approaches (see Table 2).

Table 2. Chegg-Treated Essays Designated Better in Shared Sample

Rater A	1/5	20%
Rater B	1/5	20%
Rater C	2/5	40%
Rater D	0/5	0%

The second mode of analysis assumes the raters are not interchangeable, and that each rater represents a different probability of the EasyBib Plus program's success at improving writing. Using this method, four separate samples are analyzed (n = 20), each with their own set of results. Although this method greatly lowers the sample size, 20 pairs of papers per rater maintains ecological validity in that it resembles the size of most first-year composition classes, mirroring experimental conditions to those of real-life (Brewer, 2000). The advantage of this method is that it does not assume raters all had the same approach, which the shared sample and the interviews cannot fully confirm. Given the lack of explicit direction—raters were instructed to use the deliberately-ambiguous rubric of "which is better" to assess the pairs of essays—not assuming interchangeability among raters might prove more accurate.[9] The results of both modes of analysis are provided below (see Table 3).

Using the first analytic method, the overall analysis yields a point estimate (p-hat) value of $\hat{p}$ = .30 and a 95% confidence interval (CI) of [.20, .40] for the Chegg-treated essays. In other words, out of 80 pairs of essays, raters on average designated the Chegg-treated essays better only 30% of the time (24/80); and if the experiment were repeated 100 times, we could expect 95

[9] While this method is likely more accurate, all raters share very similar backgrounds and have similar teaching and professional experiences. We can assume their enrollment in the same graduate program and similar training provides comparable assessment approaches.

of the experiments to yield a point estimate value between .2-.4, or 20-40%. The null hypothesis assumed that the treated essays would not differ meaningfully from the untreated essays in either improvement or detriment. Therefore, a null hypothesis point estimate value of $\hat{p}$ = .50 (50%) was assumed, and results were tested against this value for significance. A paired sample t-test yields a t-statistic of $t(79) = -3.90$, $p < .001$, meaning the $\hat{p}$=.30 value is statistically significant.

Analyzed independently, two raters (C, D) perceived the Chegg-treated essays better at a rate of 45% [.23, .67], slightly worse than a coin flip, which is not a statistically significant difference from the 50% probability of the null hypothesis. Conversely, the other two raters (A, B) perceived the *untreated* essays as significantly better on average, designating the Chegg-treated essays better only 10% [-.03, .23] and 20% [.02, .38] of the time, respectively, which both differ statistically significantly from 50%. Overall, the analysis puts the success of the program at approximately 30%, with the 95% Confidence Interval suggesting we could be confident that a Chegg-treated essay would on average be perceived as better only between 20-40% of the time compared to an untreated essay. Using either analysis, the AWE program at best achieves a probability of improvement of approximately 45%, roughly equivalent to a coin flip; at worst, the AWE program has significantly worse odds of improving a paper (10%) than leaving the essay unedited.

Table 3. Point Estimate Analysis of Chegg-Treated Essays Designated Better

	Treated	$\hat{p}$	SE	95% CI	*t*(df), *p*
Rater A**	2/20	.10	.07	-.03[a], .23	*t*(19)= -5.96, $p < .001$
Rater B*	4/20	.20	.09	.02, .38	*t*(19)= -3.35, $p = .0016$
Rater C	9/20	.45	.11	.23, .67	*t*(19)= -.90, $p = .19$
Rater D	9/20	.45	.11	.23, .67	*t*(19)= -.90, $p = .19$
Overall**	24/80	.30	.05	.20, .40	*t*(79)= -3.90, $p < .001$

Note. $\hat{p}$= p-hat, the probability a treated essay is designated better. SE= Standard Error. CI=Confidence Interval.

[a]The lower bound of Rater A's confidence interval is slightly negative; it is acceptable to interpret this value as 0, since the parameter it is estimating is a positive value.

Note. *$p < .01$. **$p < .001$

3.4.2 The Instructor Interviews

At the conclusion of the experiment, I conducted brief (between 5-10 minutes each) informal interviews (Blakeslee and Fleischer, 2007) with each of the raters (see Table 4). The goal of the interviews was to gain a clearer picture of how the raters approached the task of evaluating the essay pairs, which the quantitative data cannot fully reveal. If the raters varied greatly in their approach, the overall analysis (n=80) is less valid, since the difference in individual approaches would be the most determinant factor for the experimental results, and an analysis of each sample individually (n=20) would be required. However, if the raters approached the task similarly enough, they could function as interchangeable in terms of the experiment, validating the overall analysis. In addition, the interviews help to provide more

clarity about the effect of the EasyBib Plus AWE program and how the raters perceived the changes it made to the writing being evaluated.

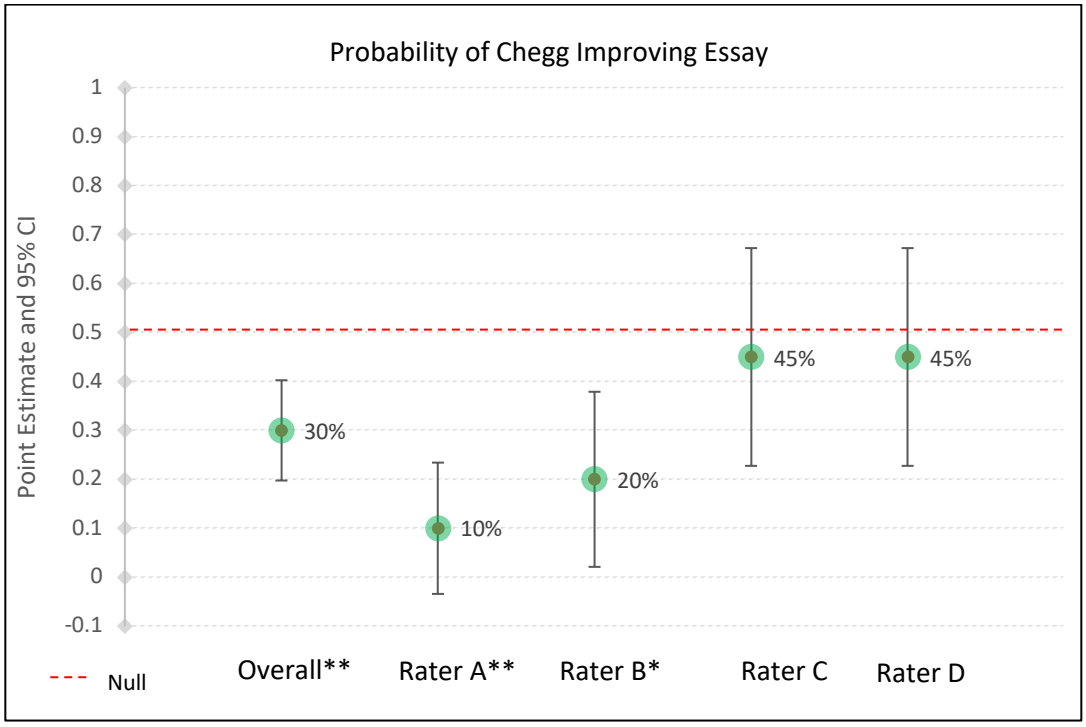
Table 4. Interview Questions

1. Basic information about instructors:
    - What year are you?
    - How many years have you taught college English classes?
    - Do you teach argumentative writing in your classes?
2. Describe the criteria you used to determine which essay was better.
3. Did your criteria change over time?
4. How confident were you in designating one essay better than the other?
5. How different did you think each pair of essays were from one another?
6. Which features of writing do you think make for a well-written editorial/argumentative essay?
7. What changes do you look for when assessing the quality of revisions from a rough to a final draft?
8. What strategies do you suggest for your students to follow when revising from a rough to final draft?

As discussed in the participants section, each of the four rater's backgrounds were extremely similar, and so their response to question one was approximately the same. Questions two and three saw some variation. Since a comparative judgment model of assessment was used, raters were free to draw on whichever criteria they saw fit in order to determine the better of the two essays per pair. For example, rater A described using criteria based on essay features such as word choice, syntax, and the overall "flow" of the language. Rater B focused their criteria on which essay generally communicated its meaning—as determined by the rater—more successfully. All raters stuck to their criteria throughout the experiment, with little change throughout.

Questions four and five prompted similar responses from all raters. Raters thought that, overall, the differences between essays in each pair were not great, which somewhat lowered their confidence in designating one of each pair better. Each of the raters described the

differences as something like "surface-level" or at the level of "lower-order" concerns. Despite the agreement that the differences were not great, each of the four raters, as well as each of their practice samples, yielded a Chegg success rate below 50%. Future research might be interested in the disconnect between these interview responses and the quantitative data yielded by the experiment—as the interview data would seem to predict each of the four raters would yield point estimates closer to the null value like those of raters C and D (see Figure 3).



Note. *$p < .01$. **$p < .001$

Figure 3. Probability of Chegg Improving Essay

Questions six, seven, and eight prompted mixed responses, but each were relatively similar. These questions were designed to spur a discussion of genre and revision pedagogy. Rater D commented on the more informal tone of editorial writing, and others echoed that sentiment. All raters discussed that the argument should be clear and persuasive in an editorial, and that effective editorials are ones that know their target audience. For revision pedagogy, rater

C made a distinction between sentence-level revision and "big picture" revision—such as that of thesis statements and entire paragraphs—and seemed to suggest that the differences in essays were primarily noticeable at the sentence level.

In sum, the interviews helped to contextualize the quantitative data yielded by the experiment. While the responses to questions about how different the essays were and the raters' confidence in designating one better than the other somewhat contradict the quantitative data, the interviews also give greater insight into the teaching of editorial writing at the college level and why, perhaps, the Chegg EasyBib Plus tool failed to improve student writing. There appears to be broad consensus on the characteristics of the genre, as well as an understanding of the difference between higher- and lower-order concerns. Nonetheless, the greater volatility in the analysis of the raters individually could reflect the variation inherent to different classes, semester by semester, since each of the raters' samples (n=20) is approximately the same size as a typical FYC course. Naturally, when several classes are combined, the variation is minimized, as the overall sample (n=80) reflects.

3.5 Discussion

The experiment found that Chegg's EasyBib Plus AWE tool was unsuccessful at improving editorial essays, as evaluated by four current college English instructors. The experiment's null hypothesis assumed the probability of designating a Chegg-treated essay "better" to be approximately 50%. In reality, on average the 80 Chegg-treated essays had only between a 20-40% probability of being designated better than their 80 untreated counterparts, a statistically significant difference from the null hypothesis value. This suggests features of the program itself, and not random chance, were responsible for rendering the treated essays worse than their unedited versions. The results should be interpreted cautiously with regards to

generalizability. Nonetheless, the results open an interesting discussion about the inherent limitations of AWE technology to formatively evaluate writing, specifically AWE's inability to parse genre and meaningfully address higher order writing concerns.

3.5.1 AWE and Genre

A major obstacle for AWE programs is genre. AWE programs have traditionally been used in summative evaluation efforts such as scoring written placement exams or standardized test essays, where test takers submit an essay written in a very narrow genre whose criteria the AWE programs have been trained on. Summative *scoring* of essays by machines correlates very highly with those of human raters (Dikli, 2006). But scoring and improving essays are very different tasks. Moreover, without the narrow genre constraints provided by a standardized essay exam, AWE programs may struggle to comprehend the larger context within which the writing has occurred, which in the case of this experiment is an argumentative editorial. The challenge of parsing genre, as well as a closer analysis of what the EasyBib Plus tool actually does, can help explain the results.

When a user logs on to Chegg to access the EasyBib Plus tool, they are prompted with a submission portal that asks for no genre information. The user uploads or copy/pastes a paper into the portal, and then the program provides suggestions for revision. There are no questions about genre or assignment guidelines; the EasyBib Plus seems to view writing as simply writing, independent of genre constraints or context. This acontextuality is likely a contributor to the tool's poor performance in the experiment. Without knowledge of the assignment's genre or context, the tool is limited to providing mostly general writing advice, which may or may not be applicable to a given genre.

For example, a common suggestion the tool made was to omit intensifier words like "very" or "just." Elimination of such words may help with an essay's formality or tone, but in the context of an editorial could temper the forcefulness, or obscure the clarity, of the argument or writer's position. The program offered many other surface-level suggestions, similarly aimed at tone or general writing "improvement" that could have conflicted with expected genre characteristics. Because I was careful to let the raters use their own criteria to define "better" however they saw fit, and since the findings were fairly consistent across four different raters, it seems reasonable to assume the program was unable to navigate the editorial genre appropriately.

3.5.2 "Teaching to the Test" and "Writing to the Program"

Another issue to consider is the flattening of 80 individual writing voices that occurs from feeding each essay through the same AWE program. Because the EasyBib Plus is insensitive to genre and therefore offers only general writing advice, many of the suggestions it made were repeated across essays. Although many of these repetitions were simple word or phrase adjustments, the effect resulted in a kind of uniform writing voice, which is likely amplified to raters reading 25 pairs of essays back-to-back.

The flattening of writing voices by AWE algorithms recalls an insight from social scientist Donald Campbell. Writing about research methodology in the 1970s, Campbell formulated what would later become known as Campbell's law, which states, "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor." In other words, Campbell is talking about the conflation of measurement

and achievement, which, for our purposes, is similar to the conflation of summative and formative evaluation mechanisms.

Campbell's law is popularly used to explain the phenomenon of "teaching to the test," which occurs when summative measures like standardized test scores are inappropriately used for accountability, through which they become corrupted and used erroneously as formative educational tools. As Campbell (1979) explains:

> Achievement tests may well be valuable indicators of general school achievement *under conditions of normal teaching aimed at general competence.* But when test scores become the goal of the teaching process, they both lose their value as indicators of educational status and distort the educational process in undesirable ways.

Just as the overvaluing of standardized test scores as evidence of academic achievement results in teachers "teaching to the test," an overreliance on AWE programs for formative writing pedagogy could result in "writing to the program." Just as students can score highly on a test without learning anything of substance beyond test-taking techniques, so too could students learn to satisfy the dictates of lower order writing conventions valued by AWE algorithms without learning anything of substance about writing itself.

3.5.3 Limitations

This experiment is limited in many ways. Although a quasi-representative and random sample of college students was obtained, 80 is a small sample size. A larger sample of student essays, as well as a larger number of raters, would provide greater confidence in the results. In addition, only one AWE program—Chegg's EasyBib Plus—is tested. Chegg is an extremely popular service, especially among college students, and while its AWE tool is comparable to similarly accessible online resources, other AWE programs might perform differently given the same experimental conditions. More research is needed about other programs, especially

regarding AWE tools like the Linguistic Inquiry and Word Count (LIWC) that claim to feature genre-sensitive analyses of writing.

A significant limitation in the study's design is the retro-treatment of student essays. Because I had to feed each essay into the EasyBib Plus tool myself and accept revision suggestions at my own discretion, rather than have the student writers perform this task themselves, the actual changes made to the essays represent an approximation of how the program might be used. I tried to be consistent in the changes I accepted, and carefully logged as many essay attributes as I could to monitor my consistency, and I tried not to deliberately accept changes that would make the program look worse, for instance. But I also accepted as many revision suggestions as I could, hoping to test the program's true abilities. In the future, research like this would be stronger if students used the AWE program themselves, to more accurately capture the variation in approaches to using the program.

3.6 Conclusion

This experiment provides evidence in support of an intuitive sense of the inherent limitations of AWE programs to meaningfully improve writing quality at the college level. Despite these intuitions, it may seem surprising that the program performed so poorly. While this experiment is limited in many ways and cannot confirm theories with total certainty, the numbers allow for informed speculation and skepticism about the abilities of AWE programs and what values underlie them. The EasyBib Plus program claims to improve grammar, clarity, and writing style so that students can turn in their "best paper," but this study suggests that using Chegg's program results in a "better" paper than an unedited first draft only 20-40% of the time. More research is needed, but in the meantime claims of AWE's formative educational value should be met with scrutiny.

CHAPTER 4: RAGE AGAINST THE MACHINES

"[to Dave] This mission is too important for me to allow you to jeopardize it."

—HAL 9000, *2001: A Space Odyssey*

4.1 Introduction

In this chapter, I offer an extended and speculative discussion about automated educational technology, particularly in a corporatized university context. While chapter three's discussion was confined to the experimental results and the narrow conclusions I could responsibly draw about the effectiveness of one particular AWE tool, this chapter discusses more broadly the relationship between online/automated pedagogy and a higher education system currently facing serious economic and public health crises.

I begin this chapter by conducting an artifact analysis of the Purdue Writing Lab and Chegg partnership. My goal is to use that case as a springboard to explore how online and automated educational technology may interface with campuses in the future, pedagogically and otherwise, and to speculate on the path forward for higher education as it exists parallel to increasingly popular online and virtual pedagogical environments. To that end, I first detail the Purdue-Chegg partnership as an instructive example of how liberal arts scholars and educators might approach the growing presence of educational technology companies on campus both prudently and politically.

I then consider the state of higher education amid the 2020 coronavirus pandemic. During this unprecedented public health crisis, nearly all colleges and universities have migrated their courses online, prompting discussions of the relative strengths and weaknesses of traditional in-person teaching and online instruction. I argue that while academic achievement and cognitive

growth are the primary purposes of colleges and universities, those outcomes are inextricably linked to the unique social and physical-campus contexts within which they transpire. The subject of writing education acutely reflects this dynamic. Online education in general, and the prospect of (semi-)automated teaching in particular, therefore threaten to disrupt higher education's ecosystem in irrevocable ways. Compounded by budget cuts, public health concerns, and questions about the value of higher education, the outsourcing of significant amounts of college learning to virtual classrooms and automated educational technology, I ultimately argue, would fundamentally mutate the DNA of higher education for the worse, and scholars and educators should fight against it when possible.

4.2 The Automated University

As the 21st century moves forward, teachers will continue to be confronted with opportunities to incorporate educational technology into classrooms and curricula. This is no truer today than it has been in the last 40 years. The difference is the degree of authority granted to current technology. Today many teachers—myself included—can hardly imagine classrooms unequipped with the internet or other "smart" pedagogical tools, and it would therefore be unwise to dismiss technology out of hand as a pedagogical *supplement*; our teaching is too mediated by technology to abandon it. The question is how to embrace technological innovation without compromising the human elements that have formed the basis of pedagogy for millennia.

Current partnerships between schools and educational technology companies offer useful cases for analyzing and anticipating the future roles technology might play in education. In the case of colleges and universities, use of educational technology in the classroom is more ambiguous than in K-12 contexts, since university administrators and departments provide much less oversight of curricula and classrooms. Individual faculty may be more or less inclined to use

varying amounts of technology; personal preferences for certain products may influence use, as might the wide range of disciplines on college campuses. Faculty in computer science departments, for instance, may have a more apparent reason to use technology than colleagues in the liberal arts.

Despite liberal arts's reputation for low-tech, humanistic pedagogy and subject matter, however, educational technology products and apps aimed at teachers and students of the liberal arts abound. As described in the history of writing assessment chronicled in chapter two, writing education, in particular, appears primed for all manner of innovative and AI-driven tools in development today. Moreover, as the definitions of writing and composition expand to include aspects of digital rhetoric such as video, audio, and photographic imagery, scholars and educators in writing studies specifically are now embracing technology traditionally reserved for technical disciplines. Ultimately, I believe this expansion is good for the field, as writing studies scholars have long anticipated a twenty-first century multiliteracy that includes a specific kind of technological acumen (The New London Group, 1996; Cope and Kalantzis, 2000; Lankshear and Knobel, 2011). Still, as scholars and educators of writing and pedagogy, we should retain a healthy and informed skepticism toward the abundance of products, especially those claiming to have formative educational faculties, even as we embrace technological advancements in other parts of the field.

4.2.1 University Inc.

It is important to remain vigilant about the motivations behind private educational technology companies and why they want to partner with universities in the first place. Textbook vendor and online college tutoring site Chegg has thus far functioned as this dissertation's test case and primary example. But college campuses are replete with corporate sponsors and

partners. From a marketing perspective, this is perfectly logical. A visible presence on campus gives educational technology companies direct access to their primary customer base, college students. Economically, it is in the material interest of such companies to forge relationships with universities, especially ones with prestigious reputations and large student populations.

The corporatization of the university has been the subject of extensive discussion throughout the 21st century. Whether the object of analysis is the increasing teaching loads of precarious adjunct instructors (Ginsberg, 2011), the declining status of the humanities and the rise of STEM (Donaghue, 2008; Nussbaum, 2011), or the steady transformation of publicly-funded universities into private, market-driven entities (Bok, 2003; Newfield, 2008; McGettigan, 2013), there is widespread consensus that universities now operate largely like businesses. The everyday rhetoric associated with higher education best reflects this corporate metamorphosis: students are now "customers" who "invest" in their education and expect a "return" in terms of a high-paying job.

Beyond the customer service model of higher education, there is a more literal corporatization taking place, as in actual corporations are setting up shop on campus. In a longform profile for *The New York Times Magazine* entitled "Why We Should Fear University Inc.," Purdue University is even presented as an example of this kind of corporatization. According to de Boer (2015), corporations are more and more taking over the physical spaces of campus, with:

> the Starbucks outpost, the Barnes & Noble as campus bookstore, the Visa card that you use to buy meals at the dining hall. Enrolling at a university today means setting yourself up in a vast array of for-profit systems that each take a little slice along the way: student loans distributed on fee-laden A.T.M. cards, college theater tickets sold to you by Ticketmaster, ludicrously expensive athletic apparel brought to you by Nike. Students are presented with a dazzling array of advertisements and offers: glasses at the campus for-profit vision center, car

> insurance through some giant financial company, spring break through a package deal offered by some multinational.

The modern-day college campus is a literal marketplace, albeit one with a quad. Crucially, de Boer's examples indicate that campus has principally corporatized the *lifestyle* domains of its students: where they eat, how they fund their tuition, the campus events they attend. The intellectual domain of the university—the classroom—remains sponsorship free, at least for now.

Despite corporate infiltration on campus being generally reserved for the domain of lifestyle markets, classrooms do not remain totally unincorporated. Total corporate infiltration of college classrooms is likely impossible; textbooks and online course hosting platforms—such as Blackboard—have typically been as far as corporations can get. But, as is the nature of the corporation, no market can remain untapped. Accordingly, I anticipate corporations will attempt to infiltrate the classroom further, especially those courses heavily mediated by technology. In fact, we are witnessing just that with the growing popularity of synchronous virtual classes (especially during the Covid pandemic), which are now commonly conducted via Zoom, a popular video conferencing application.[10] The sponsorship-free sanctity of the college classroom is something to monitor as companies continue to refine virtual and automated educational technology.

4.2.2 Liminal Pedagogical Spaces and the Corporate Backdoor

With the exceptions of textbooks, course platforms, and now online video conferencing applications, the intellectual spheres on campus that I believe are currently most susceptible to corporate penetration are classroom-adjacent. Campuses feature many spaces that host a mix of

[10] One can imagine Zoom releasing new educational tools and editions of their service aimed at educators, complete with plug-ins and apps available for purchase to "enhance" one's digital classroom.

students, teachers, and administrators, many of whom congregate for disparate purposes. Sunstein (1998) likens these spaces to "transient tables" on which we have "moveable feasts." In other words, these are liminal pedagogical spaces, like writing centers, career centers, or similar administrative sites that exist in between a traditional classroom and a student service. Scholars have long analyzed partnerships between writing centers and tech products in terms of efficacy (Neaderhiser and Wolfe, 2009), but efficacy is just one aspect. I believe "transient tables" such as writing centers and career centers will come to function as corporate backdoors into the intellectual domain of campus, and the growing corporate presence there will ultimately make classrooms more susceptible to a similar influence.

For one example, a writing center includes all the necessary ingredients for corporate creep—it does not assign grades (a summative currency reserved for university classrooms), nor does it engage in traditional classroom pedagogy with one teacher assigned to a single class for an entire semester; instead, writing centers offer individualized and formative support with a rotating cast of consultants. The one-to-one dynamic of writing centers and other liminal pedagogical spaces is therefore perfectly compatible with the corporate logic of customer-service, and is thus easier to replicate, or even automate, than other dynamics found on campus.

With advances in computational technology, companies like Chegg and LinkedIn[11] have targeted such liminal spaces, believing they can replicate—and perhaps even automate—many of their services. While Chegg's AWE tool is not threatening to replace campus writing centers anytime soon, Purdue's recent partnership with Chegg--especially given Purdue's own renowned Online Writing Lab (OWL) resource--bears examination, since it represents an example of how liberal arts scholars, administrators, and educators can work prudently and politically to establish

[11] LinkedIn uses an automated tool to help users craft their resumes, much like a consultation at a career center on campus.

a mutually beneficial relationship with such entities that preserves our standing and autonomy as writing educators and scholars.

4.2.3 Chegg and The Purdue OWL

Chegg announced a licensing partnership with the Purdue OWL in 2018. Before Chegg partnered with the OWL, another educational technology and textbook company, Pearson, had a contractual relationship with the OWL. In an interview Dr. Harry Denny, Director of the Purdue Writing Lab, explains the Pearson-OWL partnership was also primarily focused on the licensing of OWL content by Pearson in its educational materials, such as textbooks, which would link to material produced by and featured on the OWL. Although this license certainly benefited the OWL by increasing page traffic, the OWL has a longstanding reputation independent of Pearson as one of the premiere resources for writers. Therefore, Pearson likely disproportionately benefited by bolstering its reputation through an association with the Purdue OWL and by selling textbooks with linked OWL content.

Chegg similarly stands to improve its reputation through its association with the OWL. But Denny saw a chance for both the physical Purdue Writing Lab and the OWL to benefit from partnering with Chegg in a way that it did not with former partner Pearson. For one, Denny was able to negotiate the creation of three Chegg-sponsored funding lines for graduate students, independent of the English department, which enabled OWL employees to focus more exclusively on the online resource. In addition, since the partnership took effect, online traffic for the OWL is up to approximately 40 million-page views per semester, because Chegg has licensed OWL content on its own popular online student resource website. Finally, from an administrative and departmental politics perspective, Denny argues that actively working with Chegg signaled a crucial and visible willingness by the English department, which technically

does not control or own the OWL despite maintaining it, to collaborate with high level administrators. This willingness to collaborate has led to longer-term plans to move and upgrade Purdue's physical writing lab[12] as well as secure seats at the table if the College of Liberal Arts administrators have plans for the OWL in the future.

Denny successfully negotiated with Chegg and Purdue's College of Liberal Arts to help upgrade the Purdue Writing Lab facility and create new opportunities for Purdue English graduate students. But the Purdue OWL-Chegg negotiations were largely irrelevant to issues of online or automated pedagogy—Denny negotiated a mutually beneficial *administrative* partnership with Chegg. Nevertheless, because of Chegg's recent foray into AWE educational technology, the company's presence on Purdue's campus is worth closer scrutiny.

Currently, no explicit relationship between the OWL and Chegg's EasyBib Plus AWE tool exists. The tool itself is not featured on the OWL website, and the physical-space Purdue Writing Lab continues to offer in-person writing consultations. But, notably, not long after securing the Purdue partnership, Chegg sought specifically to improve its own writing tutoring capability by purchasing WriteLab, an AI-based AWE technology developed by scholars at the University of California, Berkeley, for $15M. WriteLab has since been folded into Chegg's EasyBib Plus AWE tool.

Chegg's acquisition of WriteLab signals that it is serious about the future potential of AWE technology as it relates to writing pedagogy. Unlike other AWE tools, WriteLab was developed by a team that included writing experts and teachers, led by computational linguist Matthew Ramirez and English professor Donald McQuade. WriteLab, according to Ramirez, was never intended to *replace* a teacher (Corcoran, 2018), but nonetheless much of the rhetoric

[12] Plans to move Purdue's physical writing lab have been stalled due to paused discretionary spending during the coronavirus pandemic.

surrounding the technology, especially after it was acquired by Chegg, has emphasized its unique, AI-driven algorithms such that it appears to provide a lot more than simple and rote grammar instruction.

According to a 2018 Chegg press release about the acquisition, Chegg claims that WriteLab "analyzes writers' drafts and not only offers revisions and edits, but also poses questions and suggestions, encouraging students to *learn* by making decisions about what they want to say and how they want to say it" (Chegg.com, 2018, emphasis mine). Chegg goes on to claim that WriteLab's "immediate, objective and constructive" responses "help people *learn* to write better" (emphasis mine) through the tool's ability "to analyze and provide specific feedback, suggest revisions and ask questions to help users improve their writing." Finally, Chegg is careful to note that WriteLab does all this by drawing on the "latest developments in machine learning and natural language processing" and blends "data-driven analysis with proven pedagogical principles to address specific writing features such as grammar, clarity, concision and logic." Despite Ramirez's stated reservations about not replacing teachers, Chegg projects the WriteLab technology as a highly sophisticated pedagogical resource that can deeply involve itself in student learning, much like a teacher.

Purdue has projected similar messaging surrounding Chegg's writing educational services. In a 2019 press release, Purdue's director of marketing and communications Kati Pratt touts Chegg's "AI-powered platform" as a "tool to make world class writing education more accessible" (Purdue, 2019). Like Chegg, Purdue emphasizes the potential for Chegg's tool, combined with the OWL's resources, "to teach students to become better writers," careful also to note that students will receive "immediate feedback, with deep context and rich examples."

Both Chegg and Purdue stand to benefit from such messaging, but Chegg especially, due to the Purdue OWL's unmatched reputation. As the press release reminds readers: "Purdue's OWL is globally recognized as the leading authority in writing resources, and by integrating their content with Chegg Writing tools, we will help students everywhere become better writers in school and in their professional careers." While press releases are generally understood to be promotional and not objective, the rhetoric invoked should alert writing studies scholars and educators to be diligent of how this partnership appears to the public. These public-facing documents can play an important role in persuading certain stakeholders--like parents and politicians--of the "common sense" value of educational technology while lacking much needed context.

An example of much needed context could be the findings from chapter three's experiment, which suggested that Chegg's EasyBib Plus AWE tool provided no discernible improvement to, and potentially worsened, student writing as perceived by the instructors who read two versions of student-written papers. These findings indicate significant potential limitations for AWE as a college-level pedagogical tool absent any human teacher supplement. The inability of Chegg's EasyBib Plus AWE tool to improve writing points to limitations of automated educational technology to formatively evaluate, and thus teach, students writing, which also applies to online education more generally. As online and virtual educational formats—which may contain semi- or fully-automated tools—continue to grow in popularity, educators are confronted with critical challenges. Technology does have transformative power, but the tradeoffs between face-to-face and virtual education must be weighed carefully.

4.3 The Online Education Wars

Pressure on colleges to offer more and more online course offerings has been mounting for years. Such pressure is often cloaked as an effort to increase educational access, and while greater access to information and learning can be a good thing, in reality online pedagogy in higher education is primarily motivated by savings in labor costs (Rhoads et al., 2015). In addition to reducing overhead, online courses, Rhoads et al. (2015) argue, "may contribute to deskilling faculty work," and ultimately align themselves with a "neoliberal shift in public policy" that stresses "marketization and privatization over public support for broad social programs such as public education" (p. 399). With the force of the market behind it, the push for online education will likely only increase. Many of the most significant future debates in education will revolve around its status as an in-person versus virtual experience and accumulating evidence supporting the value of its status as the former is vital in the online education wars to come.

Unfortunately, the world is currently undergoing a massive natural experiment in online versus in-person education at an unprecedented scale and without our consent. In early 2020, the world saw the emergence of the novel coronavirus, Covid-19, and a subsequent pandemic. Beginning in March in the US, all manner of public events and gatherings were postponed or canceled, including the remaining in-person school year for millions of students across the country. College campuses migrated all existing courses to "distance learning" or online formats. While the future remains unclear, it appears schools, and high-enrollment college campuses in particular, will likely face significant short-term alterations to in-class instruction as we battle the disease for the next several years.

University budget cuts spurred by projected economic recessions as well as Covid-19's impact on lowered student enrollment are likely to accelerate a process of "reforming" or

"reimagining" education that was already underway. Before the pandemic, there was pressure to integrate more online and automated instruction—at both the K-12 and higher education levels—simply because of efficiency and savings in labor costs. The pandemic now provides justification for making these changes under the guise of public health. Whereas before, "reforming" and "reimagining" education were coded language for slashing the budget, now they are necessary actions to promote public health, and any resulting budget savings simply function as a happy byproduct.

The mass migration of in-person classes to online formats has prompted extensive discussion among educators, students, and the public alike about the differences between online and in-person education (Krupnick, 2020), the affordances and limitations of the former (Boggs, 2020), and the future of the traditional classroom model (O'Donnell, 2020). While online education models are distinct from automated educational tools, it is not hard to imagine how online education might function as the first step in a logical succession to fully automated education. More, a salient similarity between the two is the reduction in human interaction, a historically essential ingredient to education that is impossible to replicate in virtual formats.

Indeed, in the future the most significant point of contention in debates about automated, online, or traditional education will consist less of disputes over the technical details of one program versus another, and instead more of whether education is fundamentally a social transaction. Those in favor of traditional in-person education will have to make the argument that an important component of a student's educational experience comes from the in-person negotiation between students and teachers and students and their peers, and that the learning of academic subjects and the accumulation of knowledge is in many ways only as powerful as the social context in which it occurs.

In addition to debates about the efficacy of online teaching, we also must consider higher education's crisis of legitimacy predicated on the rapidly increasing price of tuition. In 2019-20, the average cost of one year's tuition (not including room and board) was $10,440 (in-state) and $26,820 (out-of-state) for public schools and $36,880 for private schools (The College Board, 2019). With such massive sticker prices, parents and students alike have started questioning whether higher education is indeed worth the cost. If those same priced schools shift towards more online, and especially automated, pedagogy, it will become increasingly difficult for schools to justify the current costs of tuition.

4.4 Does a College Education Matter?

The debate about the relative value of online versus in-person education prompts a deeper, more existential question about higher education: what makes college worth attending at all? Many in the academy, such as myself, obviously have a vested interest in arguing for the value of a college education in and of itself as a public good. Yet for working parents or high school students concerned about financing their degree by taking on colossal debt, doubt persists. Why, then, according to the US Census Bureau (2016) have graduation rates steadily risen since the 1940s, so that currently almost one third of American adults has a college degree? (Figure 4)

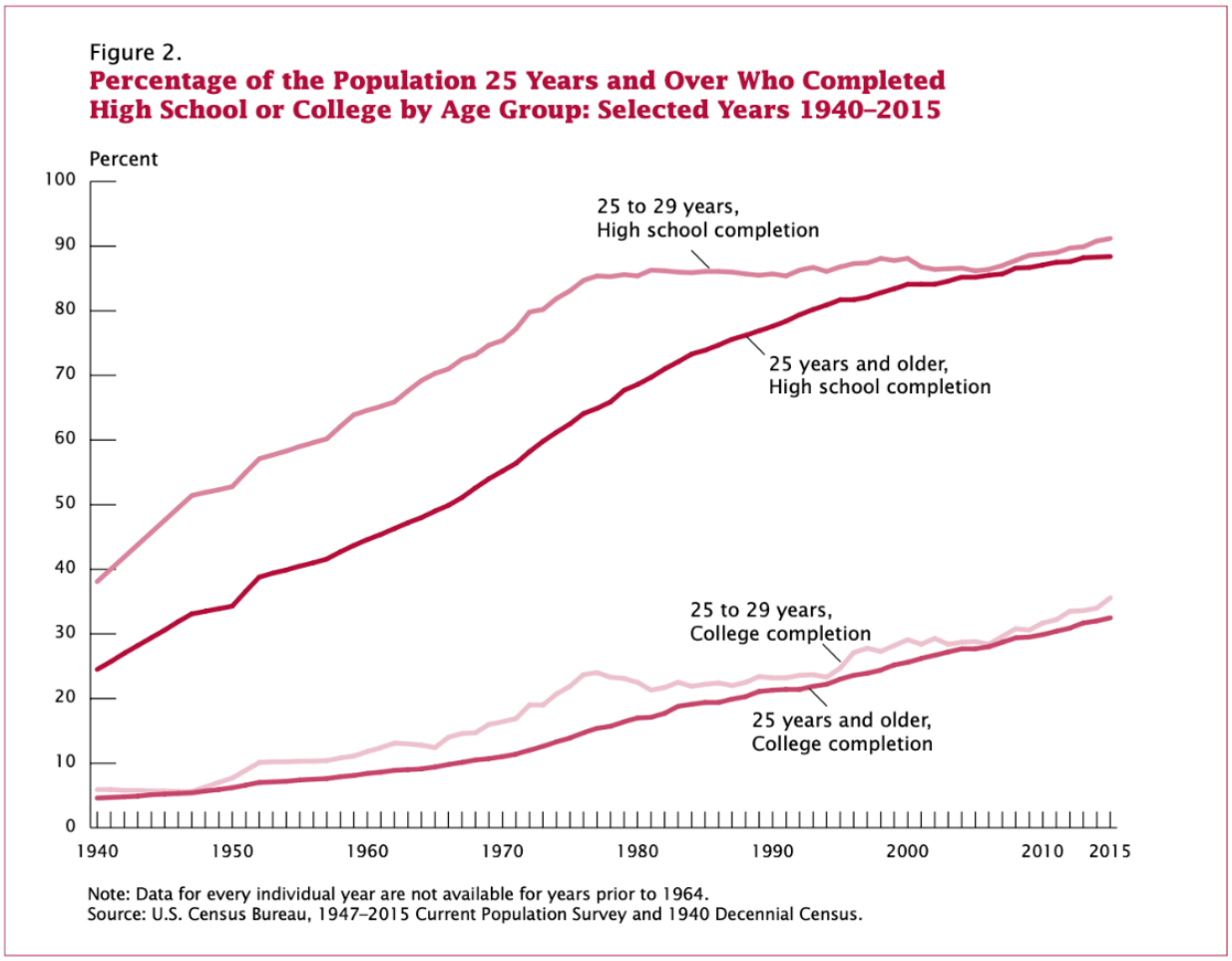


Figure 4. US High School and College Graduation Rates Over Time

One hypothesis is that college offers the only remaining path to a middle-class life—that previously available career alternatives have evaporated. But that notion is complicated by the massive amounts of debt students continue take on. The question remains: is there empirical evidence to support attending college, despite massive tuition rates? And if so, what are the benefits and why are they unique to a college education? Can we isolate a particular variable that higher education appears to impact so that we can determine what it is colleges excel at doing?

In 2014, Gallup Inc. and Purdue University partnered to conduct the most extensive survey of college graduates in history. Resulting in the Gallup-Purdue Index, the research team

sought to gather robust empirical evidence to determine the true value of a college degree beyond reductive statistics such as median salaries and job placement rates. Instead, the Gallup-Purdue Index attempts to quantify the more human elements of adult work and personal life, in order to examine the deeper value of college beyond mere assembly line-style job preparation factories. To do so, researchers focused on metrics surrounding workplace engagement, personal well-being, and emotional attachment to alma maters, which were then correlated with responses to survey questions about college experiences. With a large-*n*, representative sample size of over 30,000 college graduates, the study offers robust and key insights into what higher education is really successful at and what we as educators, scholars, and employees of institutions of higher education should work to protect. Moreover, as the next section illustrates, the research provides some of the clearest support for the importance of in-person human instruction, interaction, and mentorship to a successful college education, and the findings should caution everyone against arguments for online and automated education practices.

4.4.1 College Increases Workplace Engagement and Well-Being

The researchers were primarily interested in two variables they called "workplace engagement" and "well-being." In recognition of the fact that simply "getting any old job" may not be worth the enormous price tag of today's college tuition, the researchers identified the more nuanced variable of workplace *engagement*, which goes beyond simple statistics such as (un)employment status that can present an incomplete picture of career satisfaction. Similarly, the researchers also isolated "well-being" as a variable to understand college graduates' lives beyond their salary, which also can obscure more than it reveals. Well-being is broken down into five components: purpose well-being, financial well-being, social well-being, physical well-

being, and community well-being. Taken together, these variables attempt to reveal a truer picture of how higher education impacts adult life, both in the workplace and beyond.

The study crucially found that both workplace engagement and well-being were significantly improved by attending college (Figure 5). Critics might argue that those who attend college were already more likely to enjoy greater workplace engagement and well-being, due to the inherent privilege involved in attending college in the US. While true to an extent, another interesting finding was that *where* one attended college had no statistical difference on workplace engagement and well-being; whether one attends a public or private university, selective or comprehensive, is not predictive of engagement or well-being. By controlling for selectivity and privilege among colleges themselves, this finding suggests attending college in and of itself is at least somewhat responsible for the increase in workplace engagement and well-being, and not simply a byproduct of social privilege.

Figure 5. Impact of College Education on Workplace Engagement and Well-Being

Even more relevant are the specific aspects of higher education that contributed most to increasing workplace engagement and well-being. Researchers found that personal relationships with professors and other mentors or peers ("my professors cared about me as a person"; "I had a professor who made me excited about learning"; "I had a professor who encouraged me to pursue my goals and dreams"; "I was extremely active in extracurricular activities and organizations") increased the odds of workplace engagement and personal well-being by a factor of 1.5-2. Personal relationships and social dynamics, in other words, can lead to double the workplace engagement and well-being of college graduates, which in turn increases graduates' "satisfaction" with their college education. It is hard to imagine such social dynamics being replicated, or even imitated, in virtual classrooms.

To be sure, "satisfaction" is not the same as edification; one can theoretically be satisfied with a course—or with an entire college education—without meeting any of its intended learning outcomes. But education research increasingly suggests satisfaction and performance—as determined by course assessment instruments such as exams and papers—are associated (McFarland and Hamilton, 2005; Lee et al., 2011). Moreover, research on the differences between online and traditional classroom learning environments routinely indicate lower satisfaction rates among students enrolled in online courses and that online courses fail to recreate the social, and particularly the emotional or affective, character of in-person classroom instruction (Russo and Benson, 2005). Research detailing the limitations of online education has been well-known among individual educators and scholars for years, but massive studies like the Gallup-Purdue Index and similar projects aimed at policymakers are beginning to present widespread evidence of online education's failure to replicate face-to-face teaching (Protopsaltis and Baum, 2019).

The key takeaway from research on the relative advantages of the two pedagogical styles is that online courses overall result in significantly lower course engagement—meaning less interaction among peers and reduced communication between student and instructor, as well as less exposure to diverse others (Dumford and Miller, 2018). Depersonalized online courses eschew empirically established, effective teaching practices in favor of efficiency and cost saving at the potential expense of decreased future well-being and workplace engagement. As the Gallup-Purdue researchers conclude: "improving the college experience should focus on ways to provide students with more emotional support, and with more opportunities for deep learning experiences and real-life applications of classroom learning" (p. 23). I struggle to see how shifting to online or automated classrooms reflects this advice.

Embracing online and automated educational practices would seem to ignore the Gallup-Purdue findings, at the risk of sacrificing one of higher education's known powers. Colleges and universities offer a unique product. A college degree has traditionally enjoyed a rarefied status in the American popular imagination, but the rising cost of tuition has threatened its standing among a significant portion of the population. Add to this the increasing popularity of online course offerings and massive budget cuts, and higher education stands to lose its unique characteristics and the things it is truly good at: intellectual, *as well as social*, edification.

4.5 Higher Education's Existential Crisis

The combination of all the above—private-public corporate and university partnerships, innovative virtual educational technology, budget cuts, and public health concerns regarding current and future pandemics—presents a perfect opportunity for online and automated education advocates. Higher education was already undergoing an identity crisis in the last decade, with many media commenters declaring it "the new high school" (Vara, 2015; Selingo,

2017), and now it faces an existential crisis that educational technology opportunists surely will not let go to waste. The most valuable elements of a college education, which the Gallup-Purdue researchers have convincingly established, hinge on personal relationships among students, teachers, and their peers, are under threat.

When politicians, like New York Governor Andrew Cuomo and former Florida Governor Jeb Bush, begin calling for us to "embrace" education reform and "revolutionize" learning in the age of coronavirus (Bush, 2020; Strauss, 2020), we should be skeptical. When educational technology companies boast about their product's ability to tailor, personalize, and customize education according to each student's unique needs, we should understand that rhetoric is designed to ameliorate and disguise a dynamic that benefits neither teachers, students, nor parents.

The pandemic has simultaneously exposed the limits of technology and also created an opportunity for educational technology companies and billionaire philanthropists like Bill and Melinda Gates to peddle their products and market "interventions." All these new conditions have created a perfect storm for the further automation of education. We were headed in that direction before, but there was no explicit justification for it other than cost saving; now that we have a public health justification, attempts to automate, privatize, and disrupt public educational systems and traditional pedagogical practices will proliferate, despite reservations about technology. The "android" English teacher could be the latest educational technology product's algorithm itself, or it could be a low-paid, human "moderator" whose job is less to instruct than to robotically monitor students as they progress through various academic modules. In both cases—the algorithm and the moderator—the human element of pedagogy is eliminated.

4.6 Writing Pedagogy and The Limits of Automation

Writing is the perfect academic subject to test the limits of virtual education. To take a class in which writing is used as the primary cognitive metric for academic assessment is not to check off a list of discipline-specific knowledge; it is to interpret and embed subject knowledge in a coherent and presentable form, to grasp knowledge in such a way that students can do more than recite it—students can argue, explicate, or narrativize it. Learning through writing is not linear; it's not easily measured by successively more difficult tests that can be formulated in online modules or questions that have simple yes-no answers. Writing is by definition transactional, negotiable, and social. To write anything implies the existence of an audience—an audience of people who can read and think about and respond to the writing and thus change it.

Writing instruction is extremely tedious and complicated. It is a long-term project. It is no wonder many attempt to make the process more efficient. Outsourcing writing instruction to automated machines may feel like a solution in theory, and perhaps it may be helpful for specific purposes. But overreliance on AWE as a pedagogical resource fails to confront fundamental literacy education problems. AWE discourages human feedback and fails to sufficiently replicate the humanistic experience of school and instead devalues education by conceiving of it as something a computer module can achieve just as well as a human teacher. A writing-intensive course that relies on AWE serves as an empty gesture to the importance of education, eliminating the most important ingredient in education: human interaction, discussion, and mentorship.

Higher education classes, especially those that are writing-intensive, are not conducive to automation. These classes tend to be taught by educators with unique subject expertise and involve humanistic subjects and behaviors. The primary measure of academic performance tends to be constructed response writing, in the form of papers or essay exams. Attempts to automate or semi-automate writing assessment in such classes make sense logistically, given the

tediousness and time required to assess student writing. For large, lecture-style courses especially, where writing is less intensive and assigned more for the purpose of evaluating content knowledge, AWE technology theoretically could make such courses more efficient and save the instructors time and energy that they could devote to improving other aspects of pedagogy.

However, the discursive interaction between teacher and student through the student's writing and the teacher's feedback represents a unique knowledge-building domain in and of itself. As Emig (1977) famously described it, writing is a unique mode of learning. To automate that essential aspect of college courses is to undermine the idea that people should discuss and analyze knowledge with other people, not write about it to a computer, and modifies the very nature of higher education all together. As critics (Whithaus, 2005; Neal, 2011) have argued, the automation of writing pedagogy in liberal arts courses, for example, removes the inherent subjectivity of the content of the courses, codifying the "correctness" of the content in inflexible algorithms, which essentially transforms the courses from dialogic to dogmatic.

The debate about automating writing pedagogy specifically has familiar contours. Framed by the politics of "science versus art," the debate is most fraught at the question of what it means to teach writing. Broadly, opponents of AWE technology criticize its overemphasis on the mechanics of writing and its reduction of writing to a science (Dreschel, 1999). Instead, opponents of AWE argue, the real value of writing instruction is found in the negotiation between reader and writer. Some of these critics readily concede AWE technology's ability to consistently match essay *scores* to those of humans but argue that score consistency is not really the point of writing instruction and that rhetorical situations are contextual and thus unable to be generically automated (Baron, 1998).

Moreover, the real danger of designating a computer as the primary audience for writing is that it signifies that we no longer care about *meaningful* rhetorical situations, that writing is no longer about creating something new in dialogue with a reader but simply reproducing and reciting facts (Whithaus, 2005; Neal, 2011). The more we allow AWE to play a role in writing assessment, the more we end up letting AWE dictate how writing is to be taught and the less we determine why students should write in the first place. This recalls Garfinkle's (2020) observation in the introduction that the real danger of artificial intelligence is the prompting of humans to cede moral choices to robots and their programmers. Ceding moral choices would create a significant crisis in higher education, which thrives on the meaningful interaction between people, something machines by definition cannot replicate.

Nonetheless, AWE advocates have a point: it remains true that some aspects of writing are simply mechanical and can be aided by the computational processing power of algorithms (Breland, 1996). The question, therefore, is how much weight we give to AWE technologies in the classroom, especially as they continue to evolve to the point where they claim to be able to evaluate increasingly sophisticated aspects of writing. It is important here to distinguish instruction from identification. A program's ability to accurately identify an aspect of writing--a dependent clause, an analogy, or passive voice, for example--is not the same as its ability to instruct how those aspects can be used more or less effectively in one's writing. It is my contention that AWE technology, in its current state, is more suited for purposes of identification rather than instruction, and that its primary role should be the identification, cataloging, and analysis of written text, not the instruction of written text's production.

4.7 Conclusion

This chapter extended the discussion started in chapter three to speculate on the broader limitations and applications of AWE technology given the current conditions of higher education. As discussed, educational technology companies--and especially those selling automated teaching technology--will continue to look towards universities to target new customers and establish a presence on campus, perhaps to raise the company's profile through association. Not all partnerships between universities and private educational technology companies should be rejected out of hand; indeed, for many liberal arts educators and administrators, these partnerships could potentially function symbiotically and provide opportunities for future seats at the table to leverage the administration. We would be wise to work strategically with corporate partners, however, to try to negotiate deals that work in our favor as much as theirs. For better or worse, educational technology is not going anywhere, and we should be sure to demonstrate our own expertise to ensure technology is developed and used as properly as possible.

While I argue pedagogy is not the best use of AWE technology in particular, as supported by chapter three's findings, the coronavirus public health crisis in conjunction with budget cuts will continue to justify its widespread use. In the meantime, we would do better to identify and articulate persuasively the weaknesses of automated educational technology and virtual education and resist its incorporation into classrooms and curricula as much as we can.

CHAPTER 5: THE AGE OF AUTOMATION

> "An android," he said, "doesn't care what happens to another android. That's one of the indications we look for."
>
> —Philip K. Dick, *Do Androids Dream of Electric Sheep?*

5.1 Introduction

This chapter concludes the dissertation by summarizing the project's findings. I will begin by addressing the research questions outlined in chapter one in view of the findings and analyses presented in chapters two, three, and four. I will then direct my analysis forward to what I'm calling "the age of automation" to anticipate what awaits education generally and writing instruction specifically as the economy and American life continue to undergo automation. In addition to the earlier questions raised and answered in this project, I argue the age of automation will bring with it a mangled rhetorical landscape where it will be difficult to parse politically progressive educational policy goals from reactionary reforms to curricula and classrooms involving technology and virtual learning, and that as much as possible the core features of education should remain human-centered, despite the efficiencies and cost savings of machines.

5.2 Research Conclusions

Chapter one outlines the primary research questions and the methodology adopted to answer them. Here, I reflect on those questions:

1. *What is the history of writing assessment in the fields of rhetoric and composition and educational research? How does AWE's pivot towards formative evaluation capabilities reflect and/or disrupt writing assessment's historical trajectory?*

The history of writing assessment is not linear; instead, it oscillates. New models of assessment evolve and replace previous models based primarily on questions of validity and

reliability. The history of *automated* writing assessment parallels these oscillations. Recent questions surrounding the consequential validity of summative writing assessment, however, have prompted both writing assessment generally and automated writing assessment specifically to return to formative assessment models. As formative writing evaluation is very closely intertwined with writing pedagogy, attempts for AWE to mirror formative writing evaluation poses an unprecedented threat to in-person writing pedagogy. My analysis is that this pivot is logical but misguided—and we must continue to collect data to support claims about the limitations of AWE's formative writing evaluation efforts.

2. *What are the potential limitations and applications of AWE technology related to formative writing evaluation?*

AWE technology appears to be currently limited to the analysis and teaching of surface-level, so-called "lower-order" writing concerns. In chapter two, I discuss recent attempts by computational linguists and NLP researchers to model deeper, "higher-order" writing traits, such as strength of argument and use of non-literal language tropes like metaphors and analogies. These attempts reflect the observation that AWE developers are currently attempting to, and will continue to attempt to, refine deeper formative evaluation faculties, such as analytic feedback and the ability to assess rhetorical features of writing beyond lower-order concerns. While some of this research shows promise in its ability to *identify* these language constructions, it remains unclear that these programs can *instruct* writers in how to use them effectively.

Moreover, as detailed in chapter four's discussion about AWE and the limits of automated pedagogy, ceding the teaching of higher-order writing features, such as non-literal language tropes and argumentative writing strategies, to programmers of AWE machines represents a moral forfeiture that fundamentally transforms the teaching of writing from a negotiation between humans to a formulaic reproduction of conventions monitored by a pre-programmed

algorithm. In other words, the real limitation of AWE technology is that some of the most crucial elements of writing simply cannot be replicated by machines, because higher-order concerns are by definition unprogrammable—they are contingent and contextual, determined by human judgment according to unique rhetorical situations.

3. *Do formative AWE tools, such as Chegg's EasyBib Plus, improve the quality of college student writing? If so, how effectively?*

In chapter one, I explain that one of the most vexing questions for writing assessment scholars is establishing a clear line between writing as a product of varying quality and writing ability—does one necessarily imply the existence of the other? Yes and no. Relatedly, if a pedagogical intervention is demonstrated to improve the quality of writing, can it be said to improve the ability of the writer? Yes and no. Where summative and formative evaluation begin and end is ambiguous. Chapter three's experiment was designed to demonstrate an association between a mechanism of writing assessment—Chegg's EasyBib Plus AWE tool—and a change in the quality of writing as determined by college writing instructors. The change observed was a negative one; the EasyBib Plus did not improve the quality of writing.

Careful not to generalize too broadly, the experimental findings suggest that the EasyBib Plus tool fails to improve the quality of writing, which would seem to suggest it is similarly limited as a formative evaluation or pedagogical intervention. The experiment was designed for the EasyBib Plus to mimic what a teacher would do during the rough draft phase of an assigned paper; overall, it only resulted in a comparatively better paper than an unedited first draft 30% (20-40% Confidence Interval [95%]) of the time. This is significantly worse than the null hypothesis of 50% and suggests if the EasyBib Plus were used in place of a teacher providing feedback on rough drafts, it would be less beneficial than simply providing no feedback at all.

4. *What does the partnership between Chegg and Purdue University's OWL reflect about public-private partnerships in a corporatized American higher education system?*

I believe the Purdue-Chegg partnership is illustrative of the shape of higher education to come. Especially as colleges and universities continue to turn to technology for solutions to problems caused by public health pandemics and increasingly offer virtual courses to satisfy calls for educational access, partnerships between corporations and educational institutions will proliferate. There is a huge economic opportunity for educational technology companies to place their products between teachers and students, administrations and faculty.

While I urge general skepticism regarding these partnerships, I also believe individual cases can be negotiated in a way that is beneficial for both parties. The Purdue-Chegg partnership so far shows this. If liberal arts scholars and educators are vigilant, I believe they can leverage these partnerships in ways to benefit their own programs, as Dr. Harry Denny has shown.

5. *What aspects of writing pedagogy can be automated, if any at all? How does the reduction in human interaction—in both virtual and automated contexts—affect the teaching of writing?*

Chapter three's experiment suggests the automation of rough draft feedback has little to no beneficial effect. In the case of Chegg's EasyBib Plus specifically, the automation of feedback resulted in a lower-quality written product (approximately 70% of the time), indicating a reduction in pedagogical effectiveness and no real potential for formative evaluation as well. The retro-treatment of the essays, however, prevents the experiment from drawing conclusions about how much or how little students could have learned from using the EasyBib Plus program, so we must infer from the lower-quality of the written product as determined by the instructors.

Chapter four's extended discussion draws on the most rigorous and durable research we have about the value of a college education to support the social aspects of learning. While the

experiment involves technical limitations facing AWE as a formative evaluation tool, chapter four's discussion considers the effects of reducing the human element of not only writing instruction but education writ large. Human relationships—those between professor and student, student and peer, and the social life of campus generally—cannot be separated from the value of a college education itself. The more narrow subject of writing instruction reflects the inherent social dimension of a college education because it similarly requires human negotiation, judgment, and interaction for its successful teaching and learning.

The final two research questions will be addressed in the next sections of this chapter.

5.3 The Politics of Writing Education

6. *What political factors contribute to debates surrounding educational technology and education automation?*

The unique challenge of teaching writing serves as an instructive test case for education writ large. Writing embodies a series of paradoxes. It is both a science and an art, a technical skillset and a creative outlet. It is essential to every academic field—the medium through which scholarship transpires—yet its teaching is treated as a mere service to the university. Writing is inherently social, a communicative act between author and audience, and also a process that unfolds for long stretches in solitude. Writing is heavily mediated by technology, but also a fundamentally human endeavor. Finally, writing—and especially its teaching—is simultaneously progressive and conservative, associated with both radical expressivist disciplines and a traditional, prescriptivist and civics-oriented education. One of the main challenges for liberal arts scholars and writing educators is successfully balancing these contradictions, which become heightened in the age of automation.

An automated writing education fails to strike a balance between these divisions as human teachers do, and instead aligns with one side in each. With automated writing education, writing is *only* a science, a skill, and a nonsocial act; it is a lifeless interaction between a writer and a preprogrammed algorithm, a rote reproduction of conventions to be marked right or wrong. At root, writing is not concerned with being "right" or "wrong," but with effective communication. The automation of writing education reduces the nuance of negotiating effective communication between author and audience to a formulaic transmission of agreed-upon conventions between a word manager and a machine. What portends to happen to writing education in the age of automation is not simply a change in the way writing is taught, but a redefining of what writing is.

Whether automated writing education will come to be defined as politically progressive or conservative remains to be seen. As higher education itself undergoes a redefinition amid public health pandemics and technological progress, its political valence has grown more significant. Pew Research (Parker, 2019) has shown for years a growing partisan divide in views of higher education (Figure 6). Choice of academic majors and course content, as well as the overall value of a college degree, have become politically charged in a way they never have before. Writing occupies a peculiar dual position within this politicization: conservatives believe writing is an essential component of education and decry college students' alleged declining writing ability, yet simultaneously view the very departments that teach writing as part of a domineering leftwing culture taking over campuses.

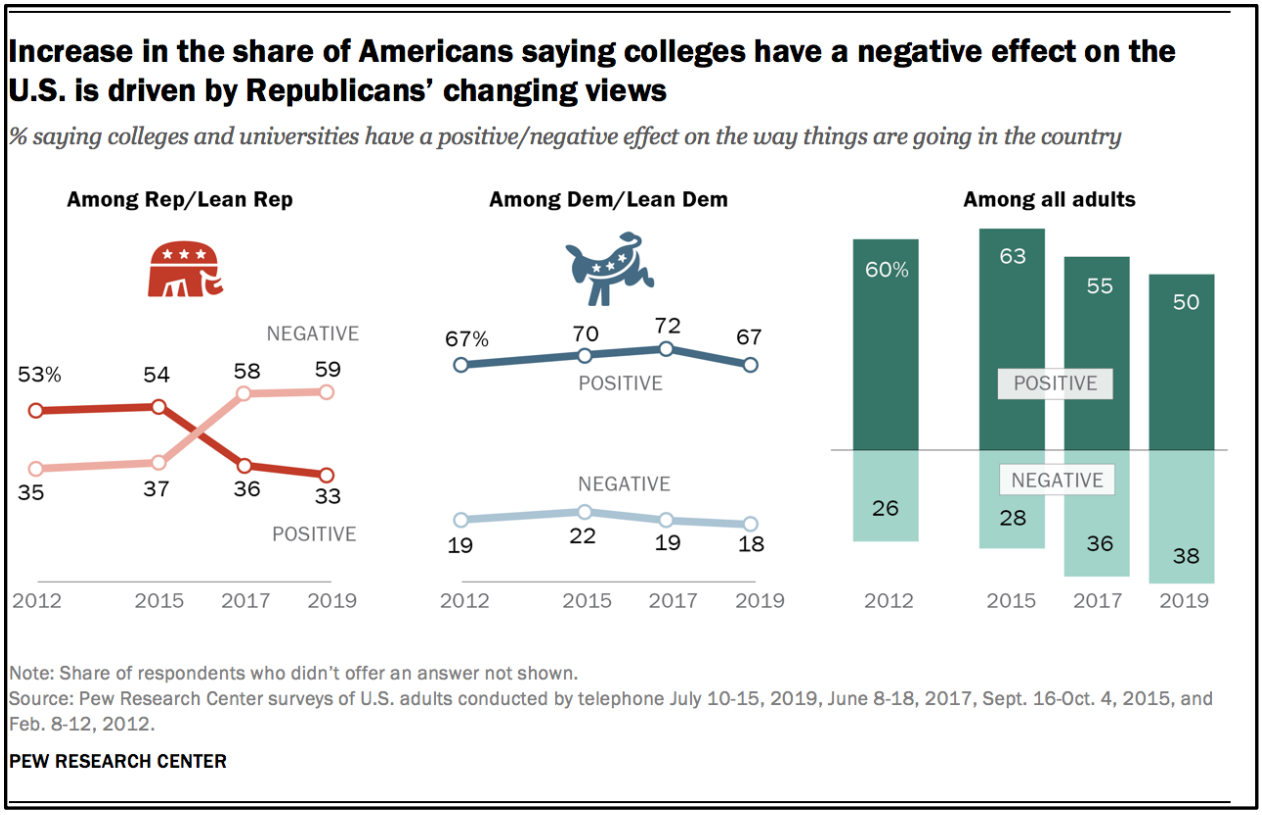


Figure 6. Growing Partisan Divide of View of Higher Education

As universities and colleges grapple with remote and virtual learning configurations in the coming years, I fear these growing political fault lines will become ammunition in those debates. Educational technology companies may enlist progressive political rhetoric to push their products, and arguments about virtual learning or automated educational technology may end up being more about political allegiances than the pedagogical effectiveness of the tools. In the event of educational technology companies—sensing an opportunity to get their foot in the door on campuses—invoking progressive political rhetoric to sell their products, we should think critically about the pedagogical repercussions of employing virtual, and potentially automated, educational products and services independent of such rhetoric as best we can.

Impressive-sounding claims of academic personalization and customization, combined with a progressive framing of pandemic-related social distancing and educational "access," will continue to escalate as education turns more and more virtual. My great fear is that out of a commanding paranoia of being perceived "conservative," liberal-minded educators will thoughtlessly accept, even advocate for, corporate-led education reforms that are nominally and symbolically progressive but deeply and structurally reactionary. Many of the arguments that preserve our autonomy as writing educators have the potential to sound conservative, and perhaps some of them even are conservative in a definitional sense. "Conservatism" has become so radioactive that people forget there are many things worth "conserving"; I believe the in-person teaching of writing is worth conserving, for instance.

We must not fear being perceived as conservative if we push back against that rhetoric. In fact, much of our jobs and livelihoods as educators revolve around conserving certain elements of the current educational model that would be irreparably disrupted by the unilateral welcoming of endless technocratic reform. If we are so afraid of being perceived as conservative that we align with nominally progressive educational reforms that beget reactionary consequences, that could give technology companies with empty progressive branding the power to significantly redefine for us what higher education looks like.

5.4 Automation All Around Us

7. *What does the increasing automation of the American economy mean for the future of education? What stands to change for students, parents, teachers, and policymakers alike?*

Automation is not a new concept; it simply looks different across historical eras. Some scholars (Guarnieri, 2010) date the first examples of automated (controlled) mechanisms as far back as ancient Greece and Egypt, noting the existence of complex time-keeping devices. The

industrial revolution offers more typical examples of automation, with factories utilizing machines to produce massive amounts of goods. In the twenty-first century, and in a country like the US which has developed a post-industrial, primarily service-based economy, contemporary automation looks a bit different still.

The automation of production-based economies is straightforward, as machines can easily replicate the mechanistic processes of assembly lines and manufacturing operations. Automating services, however, is altogether different. Service-based, or consumer-based, economies are fueled by human interaction that is harder for machines to replicate. We have seen some everyday-service components become machine automated, such as automated answering services for phone systems and automated "chatbots" or "conversational agents" on websites. These automated mechanisms assist users to perform simple tasks, such as refilling medicine prescriptions, ordering food, and booking flights. Natural Language Processing (NLP) technology is frequently used in these applications, and these programs have revolutionized aspects of the service economy.

The automation of service tasks such as these are orders of magnitude different than the "service" of education, however. Nonetheless, the automation of education is a growing topic of interest. As early as the 1980s, the prospect of computer-based automated training and education (Kearsley, 1985) has been entertained. The benefits are obvious: increased efficiency and labor cost savings. But the challenges remain, even nearly forty years later: automated educational devices lack the extemporaneous flexibility of a human instructor, the ability to adapt to specific contexts and apply human judgement to complicated situations.

Despite these challenges, a future in which education is automated is a logical response to the sociocultural context we have engineered. Veletsianos and Moe (2017) argue that the recent

"rise of educational technology" is a deeper reflection of the dominant ideologies of twenty-first century America. Since the Reagan administration, we have steadily moved away from government oversight in favor of neoliberal privatization and free-market alternatives, and the embrace of educational technology offers an impression of freedom from government-mandated curricula commensurate with that ideology. The focus on educational technology also reflects the dominance of technocracy—the idea that there is a technological fix at the root of our deepest social problems. These two paradigms combine to make education appear like any other contemporary service or product—something to be packaged and automated to better sell to customers.

If this seems farfetched or conspiratorial, consider some of the already-existing automated educational technologies of today. Writing in defense of the presence of intelligent robots in classrooms, Bushweller (2020) discusses the use of computer TAs for college courses and a robot named KeeKo used in hundreds of kindergarten classrooms in China designed to interact with students through storytelling and problem solving. In Massachusetts, an AI-powered robot named Tega helps young English language learners with literacy skills. These latter two are examples of "social robots," which are aimed precisely at replicating the latent socialization factors unique to schools. Important, too, is that these researchers have found that physical robots rather than virtual apps and modules that students interact with through screens are better for automating these social tasks—both for children and adults alike (Bushweller 2020).

As the conversation surrounding automation grows, perhaps the debate about college majors offers an instructive example to end on. The last ten years have seen abundant calls for more STEM degrees, famously exemplified by former President Obama's "STEM for all"

initiative (Handelsman and Smith, 2016). Curiously, however, even in just the last couple of years the calculus for such calls has begun to change in light of automation. Billionaire business guru Mark Cuban, for example, has recently argued, contra Obama's STEM for all, that "liberal arts is the future," precisely because the kinds of skills and knowledge gained from a liberal arts education are harder to automate than the more mechanistic knowledge found in other degrees (Jackson, 2017). This line of thinking is becoming more and more accepted, as many believe "the future workforce needs to have the skills to do the jobs that AI cannot" (Araya, 2019), which means humanity is adapting yet again to fundamental shifts in the material conditions of the labor market. If this is true, the liberal arts generally—and writing studies specifically—have potentially a big opportunity ahead of them--to offer training in a skill that is incredibly hard for machines to replicate.

If writing proves hard for machines to replicate, teaching is then impossible to automate. Beyond the loss in pedagogical effectiveness incurred by an embrace of automated pedagogy, we stand to forfeit a fundamental human relationship that has been vital to civilizations throughout history: student and teacher. If a teacher is replaced by a robot or an app or an algorithm, we stand to lose a lot more than a few points on student test averages; we lose an important dynamic through which people commune.

REFERENCES

Alkove, L. D., & McCarty, B. J. (1992). Plain talk: recognizing positivism and constructivism in practice. Action in teacher education, *14*(2), 16–22. https://doi.org/10.1080/01626620.1992.10462806

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

Araya, D. (2019). Rethinking higher education in the age of automation. *Futurithmic*. https://www.futurithmic.com/2019/06/24/rethinking-higher-education-in-age-of-automation/

Ary, D., Jacobs, L.C., Sorensen, C., & Walker, D. A. (2014). *Introduction to research in education*. Belmont, CA: Cengage.

Balfour, Stephen P. (2013). Assessing writing in MOOCs: Automated essay scoring and calibrated peer review *Research and Practice in Assessment* 8 (2013), 40-48.

Baron, Dennis. (1998). When professors get A's and machines get F's. *The Chronicle of Higher Education* (November 29).

Blakeslee, A, & Fleischer, C. (2007). *Becoming a writing researcher.* Mahwah, NJ: Lawrence Earlbaum Associates.

Bloom, B. S, Hastings, J. Thomas, Madaus, G. F, & Baldwin, T. S. (1971). *Handbook on formative and summative evaluation of student learning.* New York (N.Y.): McGraw-Hill.

Boggs, Laura. (2020). Virtual learning isn't special education. *The Wall Street Journal*. https://www.wsj.com/articles/virtual-learning-isnt-special-education-11590016215

Bok, Derek. (2003). *Universities in the Marketplace*: *The Commercialization of Higher Education*. Princeton, NJ: Princeton University Press.

Breland, Hunter M. (1996). Computer-assisted writing assessment: The politics of science versus the humanities. In Edward M. White, William D. Lutz, & Sandra Kamusikiri (Eds.) *Assessment of Writing: Politics, Policies, Practices* (pp. 249-256). New York: Modern Language Association.

Brewer, M. (2000). Research design and issues of validity. In Reis, H. and Judd, C. (eds) *Handbook of Research Methods in Social and Personality Psychology.* Cambridge: Cambridge University Press.

Bush, Jeb. (2020). It's time to embrace distance learning—and not just because of coronavirus. *The Washington Post*. https://www.washingtonpost.com/opinions/2020/05/03/jeb-bush-its-time-embrace-distance-learning-not-just-because-coronavirus/

Bushweller, K. (2020). Teachers, the robots are coming. But that's not a bad thing. *Education Week*. https://www.edweek.org/ew/articles/2020/01/08/teachers-the-robots-are-coming-but-thats.html

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning*. *2*(1): 67–90. doi:10.1016/0149-7189(79)90048-X.

CCCC. (2018). A position statement of principles and example effective practices for online writing instruction (OWI). *Conference on College Composition and Communication*. https://cccc.ncte.org/cccc/resources/positions/owiprinciples

Chegg (2018). Chegg deepens investment in writing and ai with acquisition of writelab. Retrieved August 28, 2019, from https://investor.chegg.com/Press-Releases/press-release-details/2018/Chegg-Deepens-Investment-In-Writing-And-AI-With-Acquisition-Of-WriteLab/

Comiteau, J. (2003). When does brand loyalty start? *Adweek.*com. Retrieved September 14, 2019, from https://www.adweek.com/brand-marketing/when-does-brand-loyalty-start-62841/

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

Cope, B., & Kalantzis, M. (2000). Multiliteracies: Literacy learning and the design of social futures. Psychology Press.

Corcoran, Betsy. (2018). Chegg cuts $15 million check to buy ai-feedback tool, WriteLab. *EdSurge News*. https://www.edsurge.com/news/2018-05-16-chegg-cuts-15-million-check-to-buy-ai-feedback-tool-writelab

Cotos, E. (2011). Potential of automated writing evaluation feedback. *Calico Journal*, *28*(2), 420–459.

Crossley, S., Allen, L. K., Snow, E. L., & McNamara, D. S. (2015). Pssst... Textual features... There is more to automatic essay scoring than just you! *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, 203–207. https://doi.org/10.1145/2723576.2723595

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*(1), 7–24. https://doi.org/10.1016/j.asw.2012.10.002

Deane, Paul; Frank Williams; Vincent Weng; Catherine S. Trapani. (2013). Automated essay scoring in innovative assessments of writing from sources. *Journal of Writing Assessment, 6* (1).

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, *5*(1). https://ejournals.bc.edu/ojs/index.php/jtla/article/view/1640

Donoghue, Frank (2008). *The last professor*: *The corporate university and the fate of the humanities*. New York: Fordham University Press.

Drechsel, Joanne. (1999). Writing into silence: Losing voice with writing assessment technology. *Teaching English in the Two-Year College*, 26(4), 380-387.

Dumford, A. D., & Miller, A. L. (2018). Online learning in higher education: Exploring advantages and disadvantages for engagement. *Journal of Computing in Higher Education*, *30*(3), 452–465. https://doi.org/10.1007/s12528-018-9179-z

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

Emig, Janet. (1980). The tacit tradition: the inevitability of a multi-disciplinary approach to writing research. *Reinventing the Rhetorical Tradition.* Ed. Aviva Freedman, and Ian Pringle. Ottawa: CCTE, 9-17.

Fishman, S. M. (1993). Explicating our tacit tradition: John Dewey and composition studies. College composition and communication, 44(3), 315–330. https://doi.org/10.2307/358986

Fass, D. (1991). met*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, *17*(1), 49–90.

Filliz, F. (2018). Natural language understanding. *Medium*. https://medium.com/@fahrettinf/natural-language-understanding-f50cc3229991

Gallup-Purdue. (2014). Great jobs, great lives—the 2014 Gallup-Purdue index report. *Gallup, Inc*.

Garfinkle, A. (2020). The erosion of deep literacy. *National Affairs, 43*. https://www.nationalaffairs.com/publications/detail/the-erosion-of-deep-literacy

Gedigian, M., Bryant, J., Narayanan, S., & Ciric, B. (2006). Catching metaphors. *Proceedings of the Third Workshop on Scalable Natural Language Understanding*, 41–48.

Ginsberg, Benjamin. (2011). *The fall of the faculty*: *the rise of the all-administrative university and why it matters*. New York: Oxford University Press.

Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: A meta-analysis. *The Elementary School Journal, 115(4)*, 523–547. https://doi.org/10.1086/681947

Grgurovic, M., Chapelle, C. A., & Shelley, M. C. (2013). A meta-analysis of effectiveness studies on computer technology-supported language learning. *ReCALL*, *25*, 165–198. https://doi.org/10.1017/S0958344013000013

Guarnieri, M. (2010). The roots of automation before mechatronics [historical]. *IEEE Industrial Electronics Magazine*, *4*(2), 42–43. https://doi.org/10.1109/MIE.2010.936772

Handelsman, J., and Smith, M. (2016). STEM for all. *The White House Blog*. https://obamawhitehouse.archives.gov/blog/2016/02/11/stem-all

Haswell, R. (2005). NCTE/CCCC's recent war on scholarship. *Written Communication, 22(2)*, 198-223.

Haswell, R. (2006). Automatons and automated scoring: Drudges, Black Boxes, and Dei Ex Machina. In P. F. Ericsson & R. H. Haswell (Eds.), *Machine Scoring of Student Essays* (pp. 57–78). University Press of Colorado. https://doi.org/10.2307/j.ctt4cgq0p.7

Heintz, I., Gabbard, R., Srivastava, M., Barner, D., Black, D., Friedman, M., & Weischedel, R. (2013). Automatic extraction of linguistic metaphors with lda topic modeling. *Proceedings of the First Workshop on Metaphor in NLP*, 58–66.

Herman, P. (2020). Online learning is not the future. *Inside Higher Ed*. https://www.insidehighered.com/digital-learning/views/2020/06/10/online-learning-not-future-higher-education-opinion

Hovy, D., Shrivastava, S., Jauhar, S. K., Sachan, M., Goyal, K., Li, H., Sanders, W., & Hovy, E. (2013). Identifying metaphorical word use with tree kernels. *Proceedings of the First Workshop on Metaphor in NLP*, 52–57.

Huang, T.-H. (2013). Social metaphor detection via topical analysis. *Proceedings of the IJCNLP 2013 Workshop on Natural Language Processing for Social Media (SocialNLP)*, 14–22. http://www.aclweb.org/anthology/W13-4203

Huot, B. (1996) Toward a new theory of writing assessment. *CCC, 47*, 549-567.

Huot, B. (1990). Reliability, validity, and holistic scoring: what we know and what we need to know. *College Composition and Communication*, *41*(2), 201. https://doi.org/10.2307/358160

Jackson, A. (2017). Cuban: Don't go to school for finance—liberal arts is the future. *Business Insider*. https://www.businessinsider.com/mark-cuban-liberal-arts-is-the-future-2017-2

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kearsley, G., & Seidel, R. J. (1985). Automation in training and education. *Human Factors*, *27*(1), 61–74. https://doi.org/10.1177/001872088502700106

Klebanov, B. B., Leong, B., Heilman, M., & Flor, M. (2014). Different texts, same metaphors: Unigrams and beyond. *Proceedings of the Second Workshop on Metaphor in NLP*, 11–17.

Klebanov, B. B., Leong, C. W., & Flor, M. (2015). Supervised word-level metaphor detection: Experiments with concreteness and reweighting of examples. *Proceedings of the Third Workshop on Metaphor in NLP*, 11–20.

Krupnick, Matt. (2020). Online higher education isn't winning over students forced off campus by the coronavirus. *The Hechinger Report*. https://hechingerreport.org/online-higher-education-isnt-winning-over-students-forced-off-campus-by-the-coronavirus/

Kulik, J. A., Kulik, C.-L. C., & Cohen, P. A. (1980). Effectiveness of computer-based college teaching: a meta-analysis of findings. *Review of Educational Research*, *50*(4), 525–544. https://doi.org/10.2307/1170294

Lamb, H. (2017). 'Algorithm' for metaphor development could teach computers figurative speech. Retrieved from, https://eandt.theiet.org/content/articles/2017/06/algorithm-for-metaphor-development-could-teach-computers-figurative-speech/

Landauer, T. K., Lochbaum, K. E., & Dooley, S. (2009). A new formative assessment technology for reading and writing. *Theory Into Practice*, *48*(1), 44–52. https://doi.org/10.1080/00405840802577593

Lauer, J. M. (1984). Composition studies: dappled discipline. Rhetoric Review, 3(1), 20–29. JSTOR.

Lankshear, C., & Knobel, M. (2011). *New literacies*. McGraw-Hill Education (UK).

Lee, S. J., Srinivasan, S., Trail, T., Lewis, D., & Lopez, S. (2011). Examining the relationship among student perception of support, course satisfaction, and learning outcomes in online learning. *The Internet and Higher Education*, *14*(3), 158–163. https://doi.org/10.1016/j.iheduc.2011.04.001

Mao, L., Liu, O. L., Roohr, K., Belur, V., Mulholland, M., Lee, H.-S., & Pallant, A. (2018). Validation of automated scoring for a formative assessment that employs scientific argumentation. *Educational Assessment*, *23*(2), 121–138. https://doi.org/10.1080/10627197.2018.1427570

Mason, Z. J. (2004). CorMet: A computational, corpus-based conventional metaphor extraction system. *Computational Linguistics*, *30*(1), 23–44.

MacNealy, M. S. (1999). *Strategies for empirical research in writing*. New York, NY: Longman.

Mcfarland, D., & Hamilton, D. (2005). Factors affecting student performance and satisfaction: online versus traditional course delivery. *Journal of Computer Information Systems*, *46*(2), 25-32.

McGettigan, Andrew. (2013). *The great university gamble*: *money, markets and the future of higher education*. London: Pluto Press.

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing*, *23*, 35–59. https://doi.org/10.1016/j.asw.2014.09.002

Myers, M. (2003). What can computers and AES contribute to a K-12 writing program? In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross- disciplinary perspective.* Mahwah, NJ: Lawrence Erlbaum Associates.

NCTE. (2013). NCTE position statement on machine scoring. *NCTE*. Retrieved February 21, 2018, from http://www2.ncte.org/statement/machine_scoring/

Neaderhiser, S., & Wolfe, J. (2009). Between technological endorsement and resistance: the state of online writing centers. *The Writing Center Journal*, *29*(1), 49–77. JSTOR.

Neal, Michael R. (2011). Writing assessment and the revolution in digital texts and technologies. New York: Teachers College Press.

Newfield, Christopher. (2008). *Unmaking the public university: the forty-year assault on the middle class*. Cambridge, MA: Harvard University Press.

North, S. M. (1984). The idea of a writing center. *College English*, *46*(5), 433-446. https://doi.org/10.2307/377047

Nussbaum, Martha C. (2010). *Not for profit*: *why democracy needs the humanities*. Princeton, NJ: Princeton University Press.

Parker, K. (2019). Views of higher education divided by party. *Pew Research Center's Social & Demographic Trends Project*. Retrieved June 19, 2020, from https://www.pewsocialtrends.org/essay/the-growing-partisan-divide-in-views-of-higher-education/

Perelman, L. (2014). When “the state of the art” is counting words. *Assessing Writing*, *21*, 104–111. https://doi.org/10.1016/j.asw.2014.05.001

Persing, I., & Ng, V. (2015). Modeling argument strength in student essays. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, *1*, 543–552.

Persing, I., & Ng, V. (2016). Modeling stance in student essays. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, *1*, 2174–2184.

Purdue News. (2019). The Purdue university online writing lab and Chegg partner to make world-class writing education tools more accessible. *Purdue.*edu. Retrieved August 27, 2019, from https://www.purdue.edu/newsroom/releases/2019/Q1/the-purdue-university-online-writing-lab-and-chegg-partner-to-make-world-class-writing-education-tools-more-accessible.html

Ranalli, J., Link, S., & Chukharev-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation. *Educational Psychology*, *37*(1), 8–25. https://doi.org/10.1080/01443410.2015.1136407

Reis, H. T., & Judd, C. M. (2014). Handbook of research methods in social and personality psychology (Second edition..). New York, NY : Cambridge University Press.

Rhoads, R. A., Camacho, M. S., Toven-Lindsey, B., & Lozano, J. B. (2015). The massive open online course movement, xMOOCs, and faculty labor. *The Review of Higher Education*, *38*(3), 397–424. https://doi.org/10.1353/rhe.2015.0016

Rich, C. S., & Wang, Y. (2010). Online formative assessment using automated essay scoring technology in China and U.S.—Two case studies. *2010 2nd International Conference on Education Technology and Computer*, *3*, V3-524-V3-528. https://doi.org/10.1109/ICETC.2010.5529485

Riedel, E., Dexter, S. L., Scharber, C., & Doering, A. (2006). Experimental evidence on the effectiveness of automated essay scoring in teacher education cases. *Journal of Educational Computing Research*, *35*(3), 267–287. https://doi.org/10.2190/U552-M54Q-5771-M677

Russo, T., & Benson, S. (2005). Learning with invisible others: Perceptions of online presence and their relationship to cognitive and affective learning. *Journal of Educational Technology & Society*, *8*(1), 54–62. JSTOR.

Ruth, L., & Murphy, S. (1988). *Designing writing tasks for the assessment of writing*. Greenwood Publishing Group.

Schaffhauser, B. D. (2016). Report: Education tech spending on the rise. Retrieved September 11, 2019, from THE Journal website: https://thejournal.com/articles/2016/01/19/report-education-tech-spending-on-the-rise.aspx

Schultz, K. (2006). Qualitative research on writing. *Handbook of writing research*, 357-373.

Selingo, Jeffrey. (2017). Is a college degree the new high school diploma? Here's why your degree's worth is stagnant. *The Washington Post*. https://www.washingtonpost.com/news/grade-point/wp/2017/01/13/is-a-college-degree-the-new-high-school-diploma-heres-why-your-degrees-worth-is-stagnant/

Simon, S. (2012) Robo-readers—The new teachers' helper in the U.S. *Reuters*. https://www.reuters.com/article/usa-schools-grading-idINDEE82S0GC20120329

Shermis, M. D., Burstein, J., & Leacock, C. (2006). Applications of computers in assessment and analysis of writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research.* New York: Guilford Publications.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, *20*, 53–76. https://doi.org/10.1016/j.asw.2013.04.001

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation: current applications and new directions*. New York, NY: Routledge.

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: a cross-disciplinary perspective*. New York, NY: Routledge.

Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. *International Encyclopedia of Education*, *4*(1), 20–26.

Shutova, E. (2010). Models of metaphor in NLP. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 688–697.

Smith, C. S. (2019). Dealing with bias in artificial intelligence. *The New York Times*. https://www.nytimes.com/2019/11/19/technology/artificial-intelligence-bias.html

Smith, T. (2018). More states opting to "robo-grade" student essays by computer. NPR.org. Retrieved January 26, 2020, from https://www.npr.org/2018/06/30/624373367/more-states-opting-to-robo-grade-student-essays-by-computer

Sternlicht, A. (2018). His company WriteLab was acquired by Chegg before he turned 30. *Forbes*. Retrieved September 14, 2019, from https://www.forbes.com/sites/alexandrasternlicht/2018/05/25/his-company-writelab-was-acquired-by-chegg-before-he-turned-30/

Strauss, Valerie. (2020). Cuomo questions why school buildings still exist—and says New York will work with Bill Gates to 'reimagine education.' *The Washington Post*. https://www.washingtonpost.com/education/2020/05/06/cuomo-questions-why-school-buildings-still-exist-says-new-york-will-work-with-bill-gates-reimagine-education/

Sunstein, B. S. (1998). Moveable feasts, liminal spaces: writing centers and the state of in-betweenness. *The Writing Center Journal*, 18(2), 7–26. JSTOR.

Teddlie, C., & Tashakkori, A. (2009). Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences. Los Angeles: SAGE.

The College Board. (2019). Student budgets, 2019-20. *The College Board*.

The New London Group. (1996). A pedagogy of multiliteracies: Designing social futures. Harvard educational review, 66(1), 60-93.

Vara, Vauhini. (2015). Is college the new high school? *The New Yorker*. (https://www.newyorker.com/business/currency/college-new-high-school

Veletsianos, G., and Moe, R. (2017). The rise of educational technology as a sociocultural and ideological phenomenon. *Educase Review*. https://er.educause.edu/articles/2017/4/the-rise-of-educational-technology-as-a-sociocultural-and-ideological-phenomenon

Vojak, C; Kline, S; Cope, B.; McCarthey, S.; Kalantzis, M. (2011). New spaces and old places: An analysis of writing assessment software *Computers and Composition* 28.2 (2011), 97-111.

Wang, P. (2013). Can automated writing evaluation programs help students improve their English writing? *International Journal of Applied Linguistics & English Literature*, *2*(1), 6–12. https://doi.org/10.7575/ijalel.v.2n.1p.6

Wang, P. (2015). Effects of an automated writing evaluation program: student experiences and perceptions. *Electronic Journal of Foreign Language Teaching*, *12*(1).

Ware, P. (2011). Computer-generated feedback on student writing. *TESOL Quarterly*, *45*(4), 769–774. JSTOR.

Whithaus, Carl. (2005). Teaching and evaluating writing in the age of computers and high-stakes testing. Mahwah, NJ: Lawrence Erlbaum

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

Williams, R., & Dreher, H. (2005). Formative assessment visual feedback in computer graded essays. *Journal of Issues in Informing Science and Information Technology, 2,* 23-32.

Williamson, D. M. (2009). A framework for implementing automated scoring. *Annual Meeting of the American Educational Research Association and the National Council on Measurement in Education, San Diego, CA*.

Williamson, D. M. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues & Practice*, *31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Williamson, M. M. (2003). Validity of automated scoring: prologue for a continuing discussion of machine scoring student writing. *Journal of Writing Assessment*, *1*(2), 85-104.

Wilson, J. (2017). Associated effects of automated essay evaluation software on growth in writing quality for students with and without disabilities. *Reading and Writing*, *30*(4), 691–718. https://doi.org/10.1007/s11145-016-9695-z

Yancey, K. B. (1999). Looking back as we look forward: historicizing writing assessment. *College Composition and Communication*, *50*(3), 483–503. https://doi.org/10.2307/358862

Yancey, K. B. (2012). Writing assessment in the early twenty-first century: A primer. In K. Ritter & P. K. Matsuda (Eds.), *Exploring Composition Studies* (pp. 167–187). University Press of Colorado. http://www.jstor.org.ezproxy.lib.purdue.edu/stable/j.ctt4cgjsj.13

APPENDIX A. IRB CONFIRMATION

PURDUE UNIVERSITY

HUMAN RESEARCH PROTECTION PROGRAM
INSTITUTIONAL REVIEW BOARDS

To:	WEISER, IRWIN H
From:	DICLEMENTI, JEANNIE D, Chair Social Science IRB
Date:	04/15/2019
Committee Action:(1) (4)	Determined Exempt, Category **(1) (4)**
IRB Action Date:	04 / 10 / 2019
IRB Protocol #:	1904021987
Study Title:	The Android English Teacher: Automated Writing Evaluation and the Shape of Assessment to Come

The Institutional Review Board (IRB) has reviewed the above-referenced study application and has determined that it meets the criteria for exemption under 45 CFR 46.101(b).

Before making changes to the study procedures, please submit an Amendment to ensure that the regulatory status of the study has not changed. Changes in key research personnel should also be submitted to the IRB through an amendment.

General

- To recruit from Purdue University classrooms, the instructor and all others associated with conduct of the course (e.g., teaching assistants) must not be present during announcement of the research opportunity or any recruitment activity. This may be accomplished by announcing, in advance, that class will either start later than usual or end earlier than usual so this activity may occur. It should be emphasized that attendance at the announcement and recruitment are voluntary and the student's attendance and enrollment decision will not be shared with those administering the course.
- If students earn extra credit towards their course grade through participation in a research project conducted by someone other than the course instructor(s), such as in the example above, the students participation should only be shared with the course instructor(s) at the end of the semester. Additionally, instructors who allow extra credit to be earned through participation in research must also provide an opportunity for students to earn comparable extra credit through a non-research activity requiring an amount of time and effort comparable to the research option.
- When conducting human subjects research at a non-Purdue college/university, investigators are urged to contact that institution's IRB to determine requirements for conducting research at that institution.
- When human subjects research will be conducted in schools or places of business, investigators must obtain written permission from an appropriate authority within the organization. If the written permission was not submitted with the study application at the time of IRB review (e.g., the school would not issue the letter without proof of IRB approval, etc.), the investigator must submit the

written permission to the IRB prior to engaging in the research activities (e.g., recruitment, study procedures, etc.). Submit this documentation as an FYI through Coeus. This is an institutional requirement.

Categories 2 and 3

- Surveys and questionnaires should indicate
  - only participants 18 years of age and over are eligible to participate in the research; and
  - that participation is voluntary; and
  - that any questions may be skipped; and
  - include the investigator's name and contact information.
- Investigators should explain to participants the amount of time required to participate. Additionally, they should explain to participants how confidentiality will be maintained or if it will not be maintained.
- When conducting focus group research, investigators cannot guarantee that all participants in the focus group will maintain the confidentiality of other group participants. The investigator should make participants aware of this potential for breach of confidentiality.

Category 6

- Surveys and data collection instruments should note that participation is voluntary.
- Surveys and data collection instruments should note that participants may skip any questions.
- When taste testing foods which are highly allergenic (e.g., peanuts, milk, etc.) investigators should disclose the possibility of a reaction to potential subjects.

You are required to retain a copy of this letter for your records. We appreciate your commitment towards ensuring the ethical conduct of human subjects research and wish you luck with your study.

Ernest C. Young Hall, 10th Floor - 155 S. Grant St. - West Lafayette, IN 47907-2114 - (765) 494-5942 - Fax: (765) 494-9911

APPENDIX B. EXPERIMENT PARTICIPANT CONSENT FORM

RESEARCH PARTICIPANT CONSENT FORM

Dr. Irwin Weiser
Daniel Ernst
Department of English
Purdue University
IRB No. 1904021987

Key Information

Please take time to review this information carefully. This is a research study. Your participation in this study is voluntary which means that you may choose not to participate at any time without penalty or loss of benefits to which you are otherwise entitled. You may ask questions to the researchers about the study whenever you would like. If you decide to take part in the study, you will be asked to sign this form, be sure you understand what you will do and any possible risks or benefits.

Overview

This is a study about computer-assisted writing instruction and involves your participation in two parts: an assessment experiment and a follow up one-to-one interview. We are conducting this study to better understand the value computer programs add to the teaching and improvement of student writing. Today's portion of the project will last approximately 2-3 hours. We will analyze the results over the next month.

What is the purpose of this study?

You have been asked to participate because you have experience teaching writing at the college level and you regularly read and assess student written essays. We plan to enroll you and 3 other participants with similar teaching experience in this study.

What will I do if I choose to be in this study?

- The study involves two parts, an experiment and an interview.
- In the first part, you will be given 25 pairs (50 total) of de-identified student written essays.
- Read each pair of essays.
- Designate one of each pair "better."
- Record your designations using the Qualtrics survey, the link for which will be emailed to you before the start of the experiment.
- All data will be recorded in a spreadsheet and stored on a secure drive.
- In the second part, I will interview you about your evaluation process used during the experiment.
- Questions may also involve your opinions on writing assessment generally and your teaching experience.

How long will I be in the study?

Your participation in the experiment and follow up interview today will last approximately 2-3 hours. Additional follow up interviews may be requested at your convenience.

What are the possible risks or discomforts?
Risks for the experiment and follow up interview are minimal. Possible risks are no greater than you would expect to encounter in daily life or during the performance of routine physical or psychological exams or tests. However, if at any point during the experiment or interview you feel psychological or emotional stress or exhaustion, feel free to take a break or withdraw from the project. If you wish not to answer any of the interview questions, please let the researcher know. Your identity will remain anonymous in any subsequent writing about this research through the use of an alias. Breach of confidentiality is always a risk with data, but we will take precautions to minimize this risk as described in the confidentiality section.

Are there any potential benefits?
There are no anticipated direct benefits to participants.

Are there costs to me for participation?
There are no anticipated costs to participate in this research.

Will information about me and my participation be kept confidential?
The project's research records may be reviewed by the study sponsor/funding agency, Food and Drug Administration (if FDA regulated), US DHHS Office for Human Research Protections, and by departments at Purdue University responsible for regulatory and research oversight. All records containing your identity will be saved on a secure USB drive, stored in a locked compartment. Only the PI and key personnel will have access to identifiable records and data. All data will be logged in a spreadsheet, which will be saved on a secure USB drive. Data and records will be kept for 10 years and then will be destroyed. Writing about experimental results will be published in aggregate. Interview participants will be anonymized through use of an alias for any written or published articles. The researchers cannot guarantee that the other study participants will not breach your confidentiality.

What are my rights if I take part in this study?
You do not have to participate in this research project. If you agree to participate, you may withdraw your participation at any time without penalty. If at any point you choose to withdraw from this study, any records or data you provided up to that point will be destroyed.

Who can I contact if I have questions about the study?
If you have questions, comments or concerns about this research project, you can talk to one of the researchers. Please contact Daniel Ernst, the first point of contact, at ernst9@purdue.edu or at (502) 724-3119, or Dr. Irwin Weiser at iweiser@purdue.edu. To report anonymously via Purdue's Hotline, see www.purdue.edu/hotline.

If you have questions about your rights while taking part in the study or have concerns about the treatment of research participants, please call the Human Research Protection Program at (765) 494-5942, email (irb@purdue.edu) or write to:

Human Research Protection Program - Purdue University
Ernest C. Young Hall, Room 1032
155 S. Grant St.
West Lafayette, IN 47907-2114

Documentation of Informed Consent

I have had the opportunity to read this consent form and have the research study explained. I have had the opportunity to ask questions about the research study, and my questions have been answered. I am prepared to participate in the research study described above. I will be offered a copy of this consent form after I sign it.

_____ _____

Participant's Signature Date

_____

Participant's Name

_____ _____

Researcher's Signature Date

APPENDIX C. EXPERIMENT PROCEDURES

INTRODUCTORY REMARKS:

- You will be reading argumentative essays written by Purdue English 106 students between Fall 2015-2017, an assignment I called The Editorial. The assignment asked them to choose a relevant topic and offer their own opinion in the form of a brief essay. Timely topics were encouraged, so the papers discuss things like the 2016 election, sexism/patriarchy, standardized testing, and similar. Consider this a content warning.
- The goal of the assignment was for students to practice writing about topics with no clear cut right or wrong answer, and to write persuasively about their position. They were encouraged to be entertaining and even provocative, as they were instructed editorials differed from even-handed academic writing.
- I assessed their essays not only on their ability to articulate a coherent, persuasive argument, but also their writing style. Since I explained that editorials are written primarily for large public audiences, they needed to use a writing style that would grab and hold the audience's attention and clearly communicate their argument.
- Questions?
- Instructions:
    - You have a stack of 25 pairs of essays (50 total). They are between 1.5-3 pages double spaced
        - Each pair is on top of one another; please read them in order
    - You will read each pair (they share a common letter in the code in the top left), and designate one of the two "better."
    - You should read fairly closely.
    - Be sure you enter the code correctly
    - Even if you think there is no discernible difference between the pair, you must designate one better than the other.
    - To designate one better, enter the number portion of the code into the appropriate cell on the Qualtrics survey.
        - There will be one cell for each letter at the beginning of the code; you enter the number of the better essay in the cell.