Purdue University
Purdue e-Pubs

Open Access Dissertations | Theses and Dissertations

January 2015

The CLA+ and the Two Cultures: Writing Assessment and Educational Testing

Fredrik B. Deboer
Purdue University

Follow this and additional works at: https://docs.lib.purdue.edu/open_access_dissertations

Recommended Citation
Deboer, Fredrik B., "The CLA+ and the Two Cultures: Writing Assessment and Educational Testing" (2015). *Open Access Dissertations*. 1358.
https://docs.lib.purdue.edu/open_access_dissertations/1358

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

Graduate School Form 30
Updated 1/15/2015

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Fredrik deBoer

Entitled
THE CLA+ AND THE TWO CULTURES: WRITING ASSESSMENT AND EDUCATIONAL TESTING

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Richard Johnson-Sheehan
Chair

April Ginther

Janet Alsup

Bradley Dilger

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Richard Johnson-Sheehan

Approved by: Nancy Peterson 4/20/2015

Head of the Departmental Graduate Program Date

THE CLA+ AND THE TWO CULTURES:

WRITING ASSESSMENT AND EDUCATIONAL TESTING

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Fredrik B deBoer

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2015

Purdue University

West Lafayette, Indiana

This dissertation is for my parents.

:

ACKNOWLEDGEMENTS

I am deeply indebted to, and grateful for, my family and friends for their support throughout my six years of graduate education, especially Anika, John, Hendrik, Kerste, Aeris, and Aurora; for Susan and John Beers, who have helped me in my adult life more than I can say; for all of the faculty members who have guided and taught me, especially Richard Johnson-Sheehan and April Ginther; and for my cohort members Chris Dorn, Jeff Gerding, Charlotte Hyde, Christine Masters Jach, Beth Jones, Gracemarie Mike, Stacy Nall, Fernando Sanchez, Ellery Sills, Luke Redington, and Kyle Vealey, who have made this journey one of friendship, support, and joy. I would also like to thank my haters for making me famous.

TABLE OF CONTENTS

Page

Page

ABSTRACT

deBoer, Fredrik B. Ph.D., Purdue University, May 2015. The CLA+ and the Two Cultures: Writing Assessment and Educational Testing. Major Professor: Richard Johnson-Sheehan.

The CLA+ and the Two Cultures: Writing Assessment and Educational Testing concerns the Collegiate Learning Assessment+, a standardized test of collegiate learning currently being piloted at Purdue, and its potential impacts on writing programs and pedagogy. From an empirical, theoretical, and historical perspective, I consider the test as an assessment of writing and college learning, and use it as a lens through which to understand traditional antagonisms between writing instructors and the educational testing industry. My research details the institutional and political conditions that led to the rise of the standardized assessment movement nationally and locally, and analyzes results from Purdue's piloting program for the test. I argue that literacy educators must adapt to the increasing prevalence of standardized testing at the collegiate level in a way that preserves our independence and autonomy, and that if undertaken with care this adaptation need not jeopardize our traditional ideals.

CHAPTER 1. INTRODUCTION

A Growing Movement for Change

The American university is undergoing tumultuous and often-painful changes. Long considered a key aspect of contemporary American financial success, going to college has become a riskier proposition in recent years. Tuition costs have risen rapidly in the last decade ("Average Rates"), leading to high student loan debts for many recent graduates ("Student Debt"). Coupled with the weak labor market that has been a persistent factor of the United States economy following the financial crisis of 2008, this debt represents a major burden on young Americans just beginning their adult lives. Efforts to slow this growth in tuition are hampered by a major decline in state funding for public universities (Oliff et al). Meanwhile, enrollments have skyrocketed, with overall attendance at degree-granting institutions rising 32% from 2001-2011, according to the National Center for Educational Statistics ("Fast Facts"). These financial constraints contribute to a perceived need to derive more educational value out of limited resources.

But while the current push for greater efficiency is influenced by current events, it is also part of a long evolution in the culture and structure of the university. As a generally non-profit venture, and one tied to traditional definitions of education as a method for establishing civic virtues and societal goods, the American university has

typically defined itself in ways contrary to the for-profit culture of business. But in recent decades, observers of the academy have argued that it has undergone a neoliberal or corporate turn, adopting the rhetoric and values of big business. Terms borrowed from the corporate world like "disruption," "value-added," and "synergy" have become common. While the use of these terms might themselves merely be artifacts of fads within higher education administration, there is little question that educators and administrators within the American university feel new pressure to achieve goals typically associated with the corporate world.

The Assessment Mandate

Assessment is a key part of this change. After all, the first step of asking how an organization can do better is to ask how well it is currently doing. "Accountability" is a common trope in reform efforts in the contemporary university, with productive reform often represented as a kind of morality play in which appropriately apportioned praise and blame lead inevitably to beneficial change. In part, this attitude stems from an influential government report, *A Test of Leadership: Charting the Future of U.S. Higher Education* (2006), referred to as the Spellings Commission Report or simply Spellings Report, after former Secretary of Education Margaret Spellings, who oversaw the development of the document. Although the report praised the American higher education system as a whole for its international reputation and sterling research record, it also expressed concern over the lack of information regarding student learning. As Richard Shavelson, an expert on higher education policy, writes in his book *Measuring College Learning Responsibly* (2010), "The Commission report and the multiple and

continuing responses to it set the stage for examining assessment and accountability in higher education" (4).

These concerns were amplified with the publication of Richard Arum and Joseph Roksa's high-profile book *Academically Adrift* (2011), which argues that American college students gain little in the way of applicable learning during their college careers. The fundamental mechanism through which Arum and Roksa examined college learning was a then-new, little-discussed test called the Collegiate Learning Assessment (CLA). Developed by the Council for Aid to Education (CAE), a New York-based nonprofit organization expanding its interests from tracking financial aid in higher education to providing assessments, the test had been performed at several hundred colleges by the time of the book's publication. But *Academically Adrift* brought the assessment into keen focus in a way that is quite rare for any type of standardized test. The claim of limited learning in American colleges, and the book in general, has been particularly controversial, with many questioning its methodology and its lack of transparency (see, for example, Haswell 2012). A later study by the CAE, examining a far larger number of colleges and a full freshman-to-senior academic cycle, found a considerably higher level of improvement than *Academically Adrift* ("Does College Matter?"), significant because it used the same mechanism and a dramatically larger sample size.

Still, despite this pushback, there's little question that *Academically Adrift*'s argument found its way into popular consciousness. Here at Purdue University, President Mitch Daniels has referred to the book as "his bible." It is therefore little wonder that when the Daniels administration began pursuing a standardized assessment to implement at Purdue in early 2011, the CLA's successor, the CLA+, attracted significant early

attention. The Daniels administration, controversial from its earliest days, enacted a set of sweeping reforms in its first several years. The push for a standardized assessment was seen internally as an essential element – perhaps the essential element – of those reforms. Having appointed a task force to weigh various options among the prominent tests of college learning, the task force initially strongly recommended the CLA+ as its mechanism of choice. Over time, that selection process became more complicated, as various stakeholders within the institution influenced the selection, demonstrating the complex interplay between the needs and desires of upper administration and those of the faculty. As of this writing, the future of the test at the university is unclear, but in the larger perspective, there is no question that issues of assessment, accountability, and who is ultimately responsible for measuring student learning will persist into the future.

The Role of Writing

College writing scholars and administrators have particular interests, and particular vulnerability, in this conversation, as assessment has been a historically undertheorized aspect of college writing pedagogy. As Brian Huot writes in his book *(Re)Articulating Writing Assessment* (2002), "Unfortunately, writing assessment has never been claimed as a part of the teaching writing" (1). This does not mean that writing assessments have not been commonly undertaken at the collegiate level. Rather, these assessments have typically a) been defined in terms of crisis, remediation, or deficiency, and b) generally unconnected to broader theories and pedagogies of writing. In keeping with the largely ad-hoc nature of college writing pedagogy's development, and the difficult birth of composition as a research field, much early assessment practices were cobbled together in ways that lacked rigor, a strong theoretical framework, or consistency.

Because composition was typically seen as a lesser concern for academics, and the teaching of composition a "service" role rather than a truly academic role, there was little in the way of explicit theories of writing assessment or shared notions of best practices.

Over the past several decades, a robust field of writing assessment has at last emerged. Driven by administrative requirement, pedagogical need, and research interest, scholars from within the field of writing studies have developed a broad empirical and theoretical literature concerning the study of how well students write and how best to measure that ability responsibly. "The plethora of books written by scholars within the field of Rhetoric and Composition about writing assessment over the past ten years," writes William Condon in a 2011 review essay, "is a strong indication that the conversation about writing assessment has reached a kind of tipping point" (163). A tipping point, that is, that demonstrates the degree to which research on assessment has gone from being a kind of academic grunt work to being seen as an important and valued aspect of our discipline.

But despite this growth, significant challenges remain to developing our own research on the kinds of tests that are advocated for in the current political moment. A persistent divide between the techniques and beliefs of scholars in writing studies and those of the educational testing industry dulls our ability to impact the development and implementation of such tests. Huot writes of a "lack of integration of scholarship within writing assessment," where "literature has been written and read by those within a specific field who have little or no knowledge or interest in the other approach" (25). That divide in beliefs and practices—the existence of two cultures—is a preeminent

concern of this research. Working to bridge that gap is an essential element of preserving the disciplinary standing of writing into the future.

Writing was deeply embedded in the CLA, and remains so in the CLA+. The primary task on the test involves writing an essay that integrates evidence and makes a persuasive case about a particular course of action, with the scoring rubric calling for strength in both writing effectiveness and writing mechanics. Strong writing is thus a key part of individual student and college performance on the test. Writing programs are therefore clearly implicated in the results of the CLA+ assessment and assessments like it. If the CLA+ or similar mechanisms are to become essential parts of how colleges and universities develop and maintain their national reputations, then college writing classes become a natural focal point for review. Some within the field of composition will no doubt see this as a threat, a way in which our pedagogy is removed from our control and through which standardization is enforced from above. But I see it, potentially, as an opportunity. If we can carefully articulate the limits of this kind of assessment, and insist on a rigorous skepticism about what tests like the CLA+ can and cannot measure, their implementation might represent a chance to demonstrate the importance of our subject matter and the value of our teaching. If new forms of assessment are inevitable—and, given recent political and economic realities, they likely are—it is essential that members of our field find a way to make the best of them.

Understanding the Present, Facing the Future

The changes in the university I have described are embedded in an economic and political context. Some see these transformations as a necessary change that will ensure the long-term viability of the American higher education model. As Daniels, a national

Republican politician, wrote in a letter to the Purdue community, "We have a responsibility to our students and their families to ensure that we are providing a collegiate experience that prepares them to be contributing and productive citizens in their workplaces and their communities. This is a necessity because the world—potential students, employers, taxpayers and others— is demanding evidentiary proof that today's very expensive college costs are worth the price" ("A message").

Others see this movement as an effort to capture profits from non-profit entities. As David Hursh, an associate professor of teaching and curriculum at the University of Rochester and a former elementary school teacher, writes in his book *High-Stakes Testing and the Decline of Teaching and Learning* (2008), "recent education reforms are part of a larger effort by some corporate and political leaders to transform the nature of society by repealing the social democratic policies that have guided the United States for much of the last century… in the belief that they interfere with individual liberty and the efficiency of the marketplace" (2). Both proponents and skeptics of new reforms, it seems, ascribe tremendous importance to them. As Shavelson writes, "there is a tug-of-war going on today as in the past between three forces: policy makers, 'clients' [students and their parents and governmental agencies and businesses], and colleges and universities. This tug-of-war reflects a conflict among these 'cultures'" (5).

This dissertation is an attempt to understand where the higher education assessment movement comes from and where it might go; to take an in-depth look at the CLA+ and its use as a test of collegiate learning; to investigate traditional tensions between two major forces within this tug-of-war, writing instructors and researchers on one side, the educational testing community on the other; and to provide a local history of

the contested implementation of the CLA+ at a major public university. By explaining the historical, economic, and political origins of the current higher education assessment movement, I intend to help us better recognize the true motivating factors behind this push, and perhaps equip interested parties to better respond to its challenge. By examining the CLA+ and extant research considering it, I hope to provide useful information about tests of college learning for instructors, researchers, and administrators. By locating the CLA+ in a broader context with both practitioner writing assessment and standardized writing assessments created by the educational testing community, I will consider the traditional divide between these groups, and propose ways to close that divide in a way that is mutually beneficial for both. By detailing the local history of the test at Purdue, I will demonstrate the complex institutional and political dynamics that attend this type of administrative initiative. Ultimately, the purpose of this project is to grapple with new developments in how we assess learning in the contemporary university, to better position writing scholars and programs to adapt to a new reality.

STATEMENT OF THE PROBLEM

The research gap this dissertation is intended to fill concerns the CLA+ as an assessment of writing, and the potential of tests like it to represent an opportunity or a threat to writing and English programs in American universities. Additionally, it will consider the historical roots of division and tension between writing researchers and the educational testing community, and potential ways in which these groups could become better integrated in the future.

My hypothesis is that these moves toward accountability are based more in economic and political interests than in genuine educational need; that this political

movement for more educational assessment will continue; that writing programs will be compelled to adjust their teaching and assessment methods in ways that demonstrate student learning; and that ultimately, writing studies will have the opportunity to emerge from this evolution institutionally and academically stronger than before, if our community is strategic in responding to these changes.

Research questions include

- What are the economic, political, and cultural factors that are contributing to the recent push for more evidence-based, "value added" models of assessment in post-secondary education?
- What is the history of the CLA+? How does the test function as a test of college learning and a test of writing? How does the development of these assessments reflect the dynamics that contributed to the current assessment movement? In what ways does the CLA+ satisfy the expectations of the current assessment movement? In what ways does it subvert those expectations?
- What are the traditional theoretical and empirical disagreements between writing practitioners and the educational testing community? How do these groups differ in their definition of validity and reliability? How does the CLA's mechanism conform, or fail to conform, to these definitions?
- What is the local history of the assessment initiative at Purdue University? What institutional and state factors have led to the proposed implementation of the CLA+? How do various stakeholders at Purdue feel about the proposed implementation? What are arguments in its favor? In opposition? What are some of the potential consequences of this initiative?

- What are the results of Purdue's CLA+ initial piloting efforts? How do these results compare to national averages? Does the test appear to be viable in the Purdue-specific context?
- What does the assessment effort at Purdue University tell us about the relationship between national policy initiatives and local implementation of those initiatives?
- What are potential consequences of the higher education assessment movement for writing studies? For the American university? How should writing researchers and programs adapt to these changes?

DATA & METHODS

The portion of this dissertation that involves original data collection utilizes a hybrid approach, taking advantage of several different types of information and analysis. Its primary methodological orientation is historical and journalistic, drawing on techniques common to newspaper reporting and history. Unlike most historical research, this research has been conducted largely in real time, assembling information as events have unfolded on campus. Unlike most journalism, this research is intended for an academic audience, to be read and considered by as specialized audience of scholars rather than the general public. In its hybridity, this dissertation follows a long tradition of melding history and journalism. As Marcus Wilkerson wrote as far back as 1949, "the journalist is himself the historian of the present, and the record which he puts together will, when used with critical discretion, furnish valuable source material for the scholar of the future" (Nafziger and Wilkerson 11). This research is intended to serve precisely that purpose.

Journalism's place as a legitimate research methodology is complex and contested. Journalistic methods are infrequently discussed in traditional research methodology classes or textbooks. As Stephen Lamble writes in *Australian Journalism Review*, "journalism has been perceived as an orphan child methodologically" (103). In part, this has stemmed from a seeming simplicity in journalistic methods, frequently defined by reference to the "five W questions": who, what, where, when, why. Reflecting on the paucity of specific research methodologies in the field of journalism, the mass communications professor Margaret DeFleur wrote that developing a methodology "requires that the steps used in selecting and studying a problem be described and that justifications for using particular approaches be explained" (212), a process typically foreign to the practice of journalism. As Elise Prasigian writes, "No one has yet mapped the general step-by-step procedure a journalist follows before the story is written, the research process for information that so closely resembles the scientist's research process before the study report is written" (721).

But this lack of consistent methodology does not suggest that journalism inherently lacks rigor, or that the outcome of a journalistic approach cannot be taken seriously as academic research. Journalism performs an essential role in democratic society, providing a means through which the public can evaluate leadership and respond to problems as an informed citizenry. At the heart of these efforts are the simple questions of who, what, when, where, and why. As Betty Medsger, a journalist and biographer, writes for NYU's faculty web forum *Zone for Debate*,

> The who-what-when-where-how-why questions should not be ridiculed, as they have been by some in this debate, just as innovative forms of

> criticism and commentary should not be dismissed. We should all remember that people pay a high price for asking those often complex and hated questions, simple though they may sound. Who did what, when and where they did it, how and why it happened … these are, in fact, the very essence of the most courageous acts of journalism throughout history. They require a journalist's knowledge and a journalistic understanding of the matter at hand. (1)

This dissertation's original research depends upon this kind of simple-but-powerful information collection and analysis. This simplicity in purpose is not out of keeping with the traditional practice of history. As James Startt and W. David Sloan write in their *Historical Methods in Mass Communication* (2003), "history has been primarily a humanistic study, an exploration of what people have done. It is a form of inquiry into the past that asks questions about the things people have done and elicits answers based on evidence. In that process there is a story to be told and truth to be found" (2).

In order to define my methods with appropriate rigor, I have drawn from the limited extant theoretical work on journalism as a methodology. In terms of specific research materials, I follow Alan Knight in the *Australian Journalism Review* in applying a holistic approach to potential sources of evidence. As Knight writes, "Interviews, documents, surveillance and surveys are the tools of the investigative reporter…. The best investigators during the course of their investigation may draw on all of the tools at one time or another" (Knight 49). I followed this ethic in accessing and absorbing as many different types of materials as possible for this research. In terms of goals for the

collection and presentation of this research, I follow Keith Windschuttle, quoted in Lamble as describing the responsibilities of journalists in the following way:

> First, journalism is committed to reporting the truth about what occurs in the world …. Journalism, in other words, upholds a realist view of the world and an empirical methodology. Second, the principal ethical obligations of journalists are to their readers, their listeners and their viewers. Journalists report not to please their employers or advertisers nor to serve the state or support some other cause but in order to inform their audience…. Third, in the print media, journalists should be committed to good writing. This means their writing should be clear and their grammar precise. (Windschuttle 17).

Throughout data collection I have attempted to follow these principles of empirical, fact-based data gathering and a commitment to gathering information for the good of the academic community and the Purdue University community.

IRB

This research project was submitted to Purdue's Institutional Review Board (IRB) to ensure its compliance with standard ethical research practices, under a request for Expedited Review. This IRB application is attached as Appendix A. The IRB decision was IRB Review Not Required. The notification reads as follows: "We have reviewed the above-referenced project and determined that it does not meet the definition of human subjects research as defined by 45 CFR 46. Consequently, it does not require IRB review." In other words, this project was neither approved by IRB nor deemed exempt by IRB, but rather was determined to not require IRB submission at all. This is because this research

does not attempt generalization of its findings towards human subjects. That is, rather than attempting to make generalizable claims about some identified population, this research attempts to gather information from specific individuals for the purpose of building a local history. Therefore it does not meet the IRB definition of human research. This provided me with great latitude in information gathering.

Textual/Archival

In the development of the local history of Purdue's assessment initiative, I relied on several major sources of information. A key aspect of data collection was accessing texts that detailed administrative and institutional developments regarding the assessment process. Some of these have been publicly available texts that have been widely disseminated or otherwise are available for public use. This public accessibility renders such texts outside of the IRB process. Additionally, I was given access to internal documentation that was not specifically designated for public dissemination. None of the documentation I have quoted or cited in the text of this dissertation, however, are specifically considered confidential. At times, I consulted with preliminary reports that would later be made public; only the final, public versions of these documents are specifically cited for this research.

Purdue, as a public university in the Indiana system, is subject to Indiana's IC 5-14-1.5, known as the Open Doors Law, which gives the public right of access to meetings and to inspection of memoranda and minutes, with certain restrictions. While it proved unnecessary to invoke this law in order to gain access to the texts I required, I believe the statute expands the operating definition of public texts sufficiently so that the type of

texts I required lay outside of the IRB process and demonstrates the legitimacy of my use of these documents as research materials.

Among the texts I consulted include emails from President Mitch Daniels to the Purdue community; pages concerning the CLA+ on Purdue's website; articles about the CLA+ in the press such as in the *Purdue Exponent* and the *Lafayette Journal & Courier*; assorted public university documentation detailing administrative initiatives such as the Core Curriculum and assessment initiative; reports by assorted committees and task forces that took part in the assessment initiative here at Purdue; emails shared by concerned members of the Purdue community; proposal documents from for-profit testing companies seeking to implement their instruments at Purdue; and assorted additional materials that provided relevant information for this dissertation. Due to concerns about individual privacy, no emails that were not directly addressed to me personally or were not disseminated publicly to the larger Purdue community are discussed in this document, although I did receive and read such emails from multiple members of the larger Purdue community.

Using all of these documents, I built a chronology of events that led to the CLA+'s use at Purdue, with a special focus on the Daniels administration and its various reforms, including the common core initiative and the administrative consolidation of housing, student services, and undergraduate affairs. I assessed how the administration discuss its assessment efforts, how they have justified the use of the CLA+, how they frame that justification in relation to national economic and educational trends, and the rhetoric and terminology they employ in this effort. I also considered documents demonstrating resistance from faculty, and how this resistance was represented in the

local and national press. I also use these documents to assess the potential future directions and consequences of this assessment push.

Interviews

In order to deepen my investigation of the history of the CLA+'s implementation at Purdue University, I requested interviews from those within the university who are potential stakeholders in the CLA+ process. While interviews are a common research method within the humanities and social sciences, these interviews are typically undertaken for the purpose of generalization, as mentioned above. That is, interviews generally are used for the purpose of learning about some larger population than the interviewed subjects. In this research, interview subjects were primarily contacted in order to obtain specific pieces of information that were necessary for the assembly of the local history. Therefore, I did not undertake typical interview analysis procedures such as coding or grounded theory analysis. While I was not required to by IRB, I did provide my formal interview subjects with informed consent forms, an example of which is attached as Appendix B. Unfortunately, though I contacted many members of the Purdue community for this research, a significant majority declined to be quoted in this research, perhaps out of fear of institutional reprisal or out of a desire to maintain confidentiality in administrative procedures. I was, however, able to acquire the information necessary for this research. Interview transcripts are attached as Appendix C. I also was contacted by two members of the Purdue community who were willing to provide information and be quoted, so long as I protected their anonymity and did not release any information in this document that could be used to identify them. One of these participants is a senior

administrator who works in the broad domain of undergraduate education, while the other is a tenured faculty member in the College of Engineering.

In addition to these formal interviews, I also took part in many informal conversations with a large number of members of the Purdue community. Often, these conversations occurred under the condition that I could use the information gathered therein to direct future research and pursue new lines of inquiry, but not quote or cite them in my research. Frequently, conversations occurred without clearly delineated rules for what information could or could not be used in this research. In these cases, I have erred on the side of caution, and have not quote or cited that information in this dissertation. Typically, these conversations would lead me to documentary evidence that I would then be able to cite appropriately.

CHAPTER SUMMARIES

This first chapter provides an overview of my study and establishes exigency for this project by placing it into a socioeconomic and political project. By situating my project within Purdue University, writing studies, and higher education, I argue that college educators must study tests like the CLA+ in order to respond to the unique challenges and opportunities.

Chapter Two provides an in-depth history of the higher education assessment movement. I place the recent push for standardized assessment of higher education in a historical framework, explaining the recent and historical policy initiatives that have led us to this current moment. I describe how a crisis narrative has taken root in the public conception of higher education, and how recent changes to the American economy

contribute to both this narrative and the perceived need for standardized assessment of college learning.

Chapter Three considers the CLA+, discussing its history, its assessment mechanisms, its competitors and analogs, and the extant empirical research conducted using it. I consider the test's context among other tests of secondary and post-secondary education, consider the strengths and weaknesses of its approaches to assessment, and discuss the policies and procedures that its developer enacts around its implementation. I discuss possible challenges to the validity and reliability of the instrument and the ways in which the test attempts to measure "value added."

Chapter Four uses the CLA+ and higher education assessment movement to consider the traditional cultural and epistemological divide between the field of writing studies and the field of educational testing. I provide a brief history of practitioner writing assessment, and describe the differences in how writing instructors and researchers have typically cast concepts such as validity and reliability when compared to the educational testing community. I investigate the traditional sources of this cultural divide, and detail some of the consequences, particularly in terms of the (in)ability of writing studies to influence policy arguments. I ultimately argue that the true conflict is within writing studies, regarding its long turmoil about the appropriate place of epistemology in the discipline.

Chapter Five develops a local history of the assessment effort at Purdue University, detailing the rise of the Mitch Daniels administration and its extensive controversies. I examine the selection of Daniels as Purdue president, his many reforms on campus, and the development of what would become the CLA+ assessment effort. I

interview multiple stakeholders and detail various perspectives from faculty, administrators, and other Purdue community members. I present information about the piloting efforts undertaken by the Office of Institutional Assessment as part of the assessment effort. I discuss the conflict that arose between the faculty senate and the Daniels administration over the test, and what that conflict says about higher education assessment writ large.

Chapter Six concludes the dissertation and presents my perspective on the various issues contained within it. I discuss the dangers that the current state of higher education presents to writing studies, the humanities, and the American university system itself. I claim that the lack of transparency in the development and implementation of standardized assessments undermines claims that these are accountability systems and reduce public information about high-stakes, high-expenditure systems within education. I argue that scholars in writing studies must become more conversant in the techniques of empiricism, social science, statistics, and educational testing, in order to defend our traditional values and institutional autonomy, in a hostile political and policy environment.

CHAPTER 2 THE HIGHER EDUCATION ASSESSMENT MOVEMENT

In the 20th century, as the world responded to a series of massively important international events and witnessed a great leap forward in technological and scientific innovation, American colleges and universities increasingly became the subject of national attention. Whereas they once were the purview of a small economic and social elite, these schools became increasingly democratized and increasingly perceived as a vital aspect of national greatness. In particular, the rapid and vast advances in the natural and applied sciences of the 1900s made the benefits of an educated populace more and more apparent. This expansion of higher education led to a new cottage industry of national commissions and reports, intended to gauge the effectiveness and value of college teaching and research. Often, these broader reports on collegiate learning included explicit calls for more or better assessment of student progress. All have contributed to the current effort to better understand how our colleges and students are doing, and they have done so with both a focus on international competition and with a rhetoric of crisis that contributes to a sense of exigency.

The legacy of this long history of calls for more assessment in higher education can be seen in contemporary politics. In 2012, President Barack Obama summed up the conventional wisdom in an address at the University of Michigan—Ann Arbor, saying

"we want to push more information out so consumers can make good choices, so you as consumers of higher education understand what it is that you're getting" ("Remarks by the President"). As innocuous as this statement may seem, it in fact reflects a project of enormous complexity, one certain to have drastic impact on American higher education, and one destined to invite controversy. In order to understand this current national assessment effort, of which the Collegiate Learning Assessment is a part, it's necessary to explore the history of these efforts.

Truman, Eisenhower, Kennedy: Three Reports

In the conventional story of the 20th century university, few changes were more significant than those brought about by the GI Bill. The 1944 Serviceman's Readjustment Act, known as the GI Bill, provided soldiers who had served active duty with funds for college or vocational training, among other benefits. With millions of soldiers returning from World War II, an economy suddenly growing at an explosive rate, and a new class of administrative and executive jobs that required more formal education, GI Bill funds contributed to a growth in college enrollments that swelled to unprecedented levels. Within 12 years of the bill passing, some 2.2 million soldiers had used GI Bill funds to pay for tuition at a college or university (Olson 596). What's more, the increased diversity was not merely economic, but social as well. In their article "Going to War and Going to College: Did World War II and the GI Bill Increase Educational Attainment for Returning Veterans?" John Bound and Sarah Turner write "it may be that some of the most lasting impacts of increasing college enrollment for World War II veterans are not visible in educational attainment but in the form of more subtle institutional changes that widened the pipeline to elite schools to include public school graduates and students from

a wider range of ethnic, religious, and geographic backgrounds" (809). More and more Americans, from more and more demographic groups, were going off to college, and increased attention was sure to follow.

In 1946, President Harry Truman began what would become a long tradition of presidential commissions concerned with colleges and universities. The Presidential Commission on Higher Education was tasked with "examining the functions of higher education in our democracy" ("Statement by the President Making Public a Report of the Commission on Higher Education"). One of the chief purposes of the 28-member commission was to determine how well the nascent GI Bill could be extended forward into the future, past the generation that had just returned home from war. The GI Bill now is a permanent fixture of American military and college life, but there were real concerns at the time about the long-term feasibility of the program. To this end, several of the commission's members were current or former military officials. In keeping with that military bent, the commission's report, published the following year, was deeply concerned with national service and the defense potential of our colleges and universities. "[H]igher education," the report intones gravely, "must share proportionately in the task of forging social and political defenses against obliteration" (*Higher Education for American Democracy*). An additional goal of the report was also to become a commonplace: making higher education more practically accessible to ordinary Americans. The report argues that "free and universal access to education, in terms of the interest, ability, and need of the student, must be a major goal in American education" (*Higher Education for American Democracy*). But as Susan Hannah has argued, this goal was rendered toothless by the political process, and this defeat too would become

commonplace. As Hannah writes, “the goal of equal opportunity foundered on old debates over redistribution from state to church, public to private, and rich to poor” (503). Hannah goes on to describe how similar goals were ultimately set aside, due to political resistance, in the Eisenhower and Kennedy administrations as well. Overall, the report from Truman’s commission portrays the post-war American university system in a positive light, but its constant invocations of the future safety and prosperity of the republic helped to establish the high stakes of its subject matter, which would become a commonplace in the many reports of this type that followed.

Ten years after Truman tasked his commission with assessing the state of the American university, Dwight Eisenhower did the same. His Committee on Education Beyond the High School, established in 1956, had much the same aim as that of Truman’s earlier project: to audit the current standing of post-high school education in America. Eisenhower’s committee was assembled during a period of unprecedented economic prosperity, although one which was not similarly beneficial to the women and people of color who were subject to systemic inequalities. Though the Korean War had ended only a few years before, this time period has also been considered a period of stability and national ease, particularly standing in contrast to the world war that preceded it and the cultural revolution that came after. Despite the superficially sunny times, it is in this committee’s report that the crisis rhetoric, first hinted at by Truman’s commission, becomes an explicit and continuing part of this genre. The report, published in 1957, tasks colleges and universities with grappling with “the outbreak of ideological conflict and the uprooting of old political and cultural patterns on a worldwide scale” (*Second Report* 1). It is in the invocation of ideological conflict that the real exigency of Truman’s

commission becomes clear: this is a Cold War document. The committee's report is explicit in arguing for higher education as a check against global communism. "America would be heedless," the report states, "if she closed her eyes to the dramatic strides being taken by the Soviet Union in post-high school education" (1). The immediate consequence of this language was the National Defense of Education Act, which according to Hannah "provided grants and loans to students in education and the sciences as a national defense response to Sputnik" (503).

For Eisenhower, this conflation of education and national defense was essential, as it enabled his advisors to bring education to his attention, which was not always easy. As John W. Sloan argues in "The management and decision-making style of President Eisenhower" (1990), "Eisenhower believed the two most strategic policy areas were national security and the economy and he resisted the expanding efforts to crowd his agenda with such policy issues as civil rights, federal aid to education, and social welfare" (310). Therefore, tying education and national defense together was a key move by his advisor, in that it compelled him to focus his attention on an issue he was not deeply invested in. Sloan describes this strategy as an example of Eisenhower's general decision making style, which relied on delegation and the strict hierarchy of authority, likely a holdover from his military days. "Eisenhower believed that tasks of the modern presidency could not be performed by one man," writes Sloan, but "required the cooperative interaction of generalists and specialists" (310). In this way, Eisenhower's presidency presaged the consistent use of blue ribbon panels and outside experts as proxies for American presidents when considering the state of higher education. Further, the notion that higher education has a responsibility to preserve America's advantage

over other countries as a matter of national defense would go on to be a key element of the higher education assessment and accountability literature, and remains so into the present day.

In keeping with the trend, the Kennedy administration was responsible for a report of its own. Like Eisenhower's committee, Kennedy's Task Force Committee on Education was animated in large measure by Cold War fears and the urgency that came with them. (In fact, Kennedy's desire to improve American higher education was so great, he began the task force before he was officially sworn into office.) Begun in 1960, the task force operated at a time when the United States was gripped by anxiety inspired by Sputnik and the Soviet Union's lead in the space race. Like many such reports, it paid special attention to "strengthen[ing] American science and technology" and "increas[ing]…the national defense" ("Text" 6). Led by Purdue University president Frederick Hovde, the commission matched that sense of urgency with an outlandishly bold proposal, requesting $9 billion dollars for expansion of the country's colleges and universities, at a time when that figure was much larger than it is today, due to inflation. The audacity of a request of this size makes sense when Kennedy's broader political inclinations are considered. As historian Michael Meagher has argued in his 1997 article "'In An Atmosphere of National Peril': The Development of John F. Kennedy's World View," most of Kennedy's policies demonstrate his "conviction that the 1960s would represent a critical period in world history," where "the international balance of power would shift in favor of the Soviet Union, and American policy had to reflect the new conditions" (471). When viewed through that lens, the scope of Kennedy's proposed

investment in higher education makes sense. The proposal put forward by his panel matched the exigency he saw in combating the Soviet Union.

In part, the perceived need for dramatic expansion of college carrying capacity reflected the imminent arrival of the Baby Boomers into college, a massive generation that threatened to simply overwhelm the existing college infrastructure. The post-war period had brought about the Baby Boom, a demographic explosion that would provide, after a couple decades, a large new cohort of potential students. Exacerbating this trend was the Vietnam War that raged in Southeast Asia. Desperate to avoid the draft, many young men pursued college degrees to earn deferments from local draft boards. As David Card and Thomas Lemieux have documented, this resulted in significant increases in college attendance, as "the college entry rate of young men rose from 54 percent in 1963 to 62 percent in 1968 (the peak year of the draft)" (97). Swelling student populations and the increasing economic opportunity to attend college, brought about by mid-century American prosperity, contributed to the transformation of a college education from the purview of the elite to a still somewhat-rare but mass phenomenon. With more students and greater reliance on public funds came more scrutiny. This increased scrutiny was highlighted by the Kennedy commission's report being published in *The New York Times*, then as now the most prominent and influential newspaper in the country. The full requests of the task force's requests would never be met, and Kennedy was assassinated before he could see the full impact of his recommendations for expanding higher learning and research. But the task force was considered a major part of the successful passage of the Higher Education Act of 1965, signed into law by Kennedy's successor, Lyndon B. Johnson. That bill was responsible for the establishment of Pell Grants and a dramatic

expansion in the monetary availability of a college education to more and more Americans.

Although these commissions and their reports may seem remote from the current push for greater accountability in collegiate education, given their age, they have contributed to that effort by establishing a tradition of top-down investigations of the quality of our higher education system—and a tradition of crisis rhetoric that is a commonplace in this conversation. While the commissions of Truman, Eisenhower, and Kennedy did little to specify definitive assessment policies or procedures, they were essential in laying the groundwork for the calls for reform that would come next—calls for reform that were more political, more critical of our colleges and universities, and more committed to standardized assessments than ever before.

A Nation at Risk

Of the many commissions, publications, speeches, and policy initiatives that have contributed to the assessment movement, perhaps none has been more successful in causing alarm about the current state of higher education than *A Nation at Risk*, the 1983 report commissioned during the Ronald Reagan administration. A comprehensive report on the state of American education from kindergarten through college, *A Nation at Risk* was as alarmist as its title. "[T]he educational foundations of our society," reads the report, "are presently being eroded by a rising tide of mediocrity that threatens our very future as a Nation and a people" (*A Nation*). This sharp criticism of schools was in keeping with the administration's general attitudes; Reagan would go on to give fifty-one speeches advocating major school reform in his 1984 re-election campaign (Ansari). A crisp 36 pages, written by an 18-member panel working under the auspices of the

National Commission on Excellence in Education, the report was represented as a letter to the American people. Some sections are written in the second person, addressing readers as "To Parents" or "To Students" (35). Echoing the Kennedy-era task force report, *A Nation at Risk* was a highly public affair, with excerpts appearing in many newspapers and printed copies distributed widely. Eventually, some 6 million paper copies would be distributed (Guthrie and Springer 11).

The report was critical of American schooling across the age range, describing an international community rapidly closing the gap with American education and complacent schools and colleges that lacked the rigor to maintain their lead on the rest of the world. Ultimately, the commission made 38 major observations, in 5 major categories: Content, Standards and Expectations, Time, Teaching, Leadership and Fiscal Support (*A Nation*). Of particular interest to this project are those criticisms of secondary and post-secondary institutions. The report speaks of a decline in standards, high schools for failing to prepare students for college and colleges for failing to adequately respond to these deficiencies. "[T]he *average graduate* of our schools and colleges today," it reads, "is not as well-educated today as the average graduate of 25 to 35 years ago" (13, emphasis original). In a claim that will be familiar to anyone with exposure to the rhetoric of education reform, the report places the blame squarely on a lack of high standards in high school and college, rather than on economic, structural, or demographic factors. Of particular interest to this project, the report mentions a lack of rigorous graduation requirements for high school graduation and grade inflation in college as dangers to America's educational competitiveness (*A Nation* 18-19). The report advocates for higher homework loads in high school, stricter graduation requirements for seniors, and

higher minimum course, GPA, and test score requirements for college admission, perhaps under the theory that a higher minimum threshold for college acceptance would put pressure on high schools to improve student learning. Conspicuously lacking from the report is practical suggestions for how individual institutions could improve student outcomes in order to meet these more rigorous standards.

The report's criticisms spoke to a general unease about the country and its economy, owing in part to the rise of Japan and Germany as major business competitors with the United States. Though now firm geopolitical allies against the Soviet Union, these two countries had been our antagonists in World War II, still a fresh memory for many Americans of the period. "[T]he U.S. economy was tanking," writes Tamim Ansari in a 2007 retrospective on *A Nation at Risk*, "[a]nd it wasn't our enemies driving our industries into the ground, but rather our allies, Japan and Germany" (Ansari). In that sort of environment, the question of whether America's schools were producing the best young workers they could became a matter of national attention. Americans increasingly "perceived high school preparation as deficient," write Stark and Lattuca, and "wonder[ed] if and how colleges were dealing with deficiencies" (98). *A Nation at Risk* contributed to that perception. In the highly-politicized atmosphere of the early Reagan administration, *A Nation at Risk* caused considerable controversy, particularly in the context of the usually sleepy genre of federal commission reports. The response was immediate; the media "fell on the report like a pack of hungry dogs" (Ansari). As Stark and Lattuca write, "Calls for accountability in higher education at the state level quickly followed" (98).

The most prominent of those calls took the form of *Time for Results*. Though it cannot be properly thought of as a response paper to *A Nation at Risk*, given from a consortium of state governments under the directive of their governors. Part agreement with the commission's findings, part damage control, the report was an attempt by the state governments that directed and partially funded public colleges to match the rhetorical urgency of *A Nation at Risk* and to articulate a way in which to match its challenge. Published in 1991 after years of development, under the auspices of the National Governor's Association, *Time for Results* attempted to respond to the challenges of *A Nation at Risk* from the perspectives of the state governments that played and play such a large role in the development of educational policy. Lamar Alexander, then the governor of Tennessee and the president of the NGA, summarized the recommendations of *Time for Results* as follows:

- Now is the time to work out a fair, affordable career ladder salary system that recognizes real differences in function, competence, and performance of teachers.
- States should create leadership programs for school leaders.
- Parents should have more choice in the public schools their children attend.
- The nation, the states, and school districts all need better report cards about results - about what students know and can do.
- School districts and schools that don't make the grade should be declared operationally bankrupt, taken over by the states, and reorganized.

- It makes no sense, while U.S. students are undereducated and overcrowded, to keep closed for half the year the school buildings in which America has invested a quarter of a trillion dollars.
- States should work with 4- and 5 year-olds from poor families to help them get ready for school and to decrease the chances that they will drop out later.
- Better use of technologies through proper planning and training for the use of videodiscs, computers, and robots is an important way to give teachers more time to teach.
- States should insist that colleges assess what students actually learn while in college. (Alexander 202)

With the exception of using school buildings year round, this list would be familiar to anyone with knowledge of contemporary education reform movements. Most of these suggestions are common: merit pay, school choice (which means charter schools and/or private school vouchers), and increased capacity to close schools and fire teachers. Whether or not these are reasonable or responsible solutions to the problems articulated in *A Nation at Risk* is a matter of political and pedagogical debate. But this tendency of such commissions and reports to lead to findings that are consistent with the political presumptions of the administrations and organizations that commission them—and Alexander and the majority of the NGA governors responsible for *Time for Results* were conservative Republicans—calls their claims to objectivity into question. Indeed, in his overview of the report's findings, Alexander wrote that "The governors are ready for

some old-fashioned horse trading" (202), an invocation of conventional political terminology that shows the way in which partisan politics seep into ostensibly bipartisan, apolitical reports on education. The criticism of *A Nation at Risk* would further reveal this tendency.

Time for Results was far from the only response to *A Nation at Risk.* Criticism arose as well. One of the most damning set of criticisms was levied by then Secretary of Energy Admiral James Watkins and Sandia Laboratories, a set of laboratories funded by the Department of Energy—and thus, significantly, outside of the purview of the Department of Education. Watkins tasked Sandia with giving the empirical claims of *A Nation of Risk* outside review. Their report, *Education at Risk* (1991), found that in the Reagan-era commission had made significant errors in its presentation and interpretation of the then-current state of American education. In contrast with the earlier report's findings of widespread declines in various measures of educational achievement. "To our surprise," reads the report, "on nearly every measure, we found steady or slightly improving trends" (*Education at Risk*). For example, despite the claims of falling SAT scores, Sandia Labs found that no demographic subgroup of test takers had seen their average scores decline. Altogether, *Education at Risk* represented a major challenge to the crisis narrative of *A Nation at Risk.* And yet few Americans ever read it. The report was never released by the federal government, only eventually being published in a small educational journal years after its writing. The reason for this lack of publicity, it's been alleged, were political: the George HW Bush presidential campaign was running hard on notion of an education crisis, and the contrary evidence within *Education at Risk* was potentially too costly to that effort (Ansari). Whatever the reasons for the government's

refusal to publicize it, Admiral Watkins's report never attracted nearly the attention of *A Nation at Risk*. The crisis narrative firmly took hold.

This divide between the Sandia report and the earlier report it critiqued, and the disparity in the attention each received, illustrates one of the greater fears about this genre of this kind of document: they are commissioned by partisans who are looking to find a particular result for political or self-interested reasons, rather than pursuing the truth. That possibility is illustrated by a telling anecdote about *A Nation at Risk* and the president who championed it.

> As commission member Gerald Holton recalls, Reagan thanked the commissioners at a White House ceremony for endorsing school prayer, vouchers, and the elimination of the Department of Education. In fact, the newly printed blue-cover report never mentioned these pet passions of the president. "The one important reader of the report had apparently not read it after all," Holton said. (Ansari)

Response From Accreditation Agencies

While the publicity (and notoriety) engendered by *A Nation at Risk* was considerable, the immediate policy changes were less severe. This lack of immediate change in actual institutions is common to these commissions, and likely reflects on the tangled web of authority, leadership, and bureaucratic organization that dictates higher education policy in particular states and at particular institutions. The most direct changes, and the most immediate, occurred in the college accrediting agencies. In her 2002 book chapter "Accreditation and the Scholarship of Assessment," Barbara Wright argues that America's six collegiate accreditation associations were directly inspired by reform

initiatives like *A Nation at Risk* to implement a new focus on assessment. "[A]ccreditation," argues Wright, "has had a significant effect on the evolution of assessment" (240) since the major wave of calls for reform inspired by *A Nation at Risk*. "The explosive growth of the assessment movement since 1985," notes Wright, "had forced all the regional accreditation organizations to revise their procedures and place greater emphasis on assessment as a form of institutional accountability" (242). From a certain perspective, the interest accreditation agencies took in leading the assessment charge was natural; the agencies have the explicit mandate of ensuring that colleges and universities are undertaking their educational missions effectively. But as will prove a recurring theme, the real question is not whether to assess but how. "The real question," writes Wright, "is whether the linkage [of accreditation and assessment] has contributed on both sides not merely to increase *practice of* assessment but also to increasingly sophisticated *thinking about* assessment" (242, emphasis original).

In keeping with the institutional inertia that is common to such efforts, major changes to accrediting agency policies were slow to result in widespread change on the institutional level. But by the mid-90s, Wright argues, accreditation agencies had become "the most powerful contributor to assessment's staying power" (253). Wright quotes Ralph Wolff, then associate executive director of the Western Association of Schools Colleges Commission for Senior Colleges and Universities, as arguing that by 1990 accrediting agencies faced such pressure to pursue assessment measures of student learning as to essentially have no choice but to comply. Wright discusses these changes at length, including the founding of the federal agency Council for Higher Education Accreditation in 2001; changes implemented by individual accreditation agencies, such

as the Western Association of Schools and Colleges, the North Central Association, and the Southern Association of Schools and Colleges; and the beginning of the Academic Quality Improvement Project, a three-year effort to better integrate the efforts of assessment and accreditation stakeholders. Available evidence indicates that the colleges and universities have felt these changes keenly. A 1999 survey of almost 1,400 colleges and universities sought information on assessment and accreditation practices. Among the principle findings of this research was that pressure from accrediting agencies was perceived as the single greatest reason for implementing new assessment practices (Peterson et al. 8).

It's clear, then, that the call made in *A Nation at Risk* was heard by many within the broad world of American education—within the media, by politicians and state governments, by accreditation agencies, and by individual institutions. Yet while changes were made in response to the report, these changes were diffuse and inconsistent, owing to the diversity of actors involved in the process. As is typical of the federalized American system, a constant negotiation is occurring between the control of the national government, the state governments, and individual institutions. The regional accrediting agencies are, well, regional, and their response was as diverse as the parts of the country they have jurisdiction over. What remained unchanging following the publication of *A Nation at Risk* was the lack of a truly national set of recommendations and policy fixes. The next major educational commission was an attempt to create such standards.

The Spellings Commission

No event has had a more direct impact on the current collegiate assessment movement than the Commission on the Future of Higher Education, referred to as the

Spellings Commission, and its report. Given that the report was commissioned on September 19th, 2005 and released on September 26th, 2006, its relatively recent release plays a major role in this preeminence. But the Spellings commission was also uniquely responsible for the current assessment push in higher education thanks to the way it consistently identifies a lack of accountability as a key challenge to American universities, and its vocal endorsement of standardized assessments of college learning.

Spearheaded by former US Secretary of Education Margaret Spellings, for whom it is colloquially named, the commission took as its task identifying the challenges that faced the American higher education system in the 21st century. Made up of nineteen members, the commission included not only leaders from universities but also from industry, such as the CEO of the test-prep firm Kaplan and a representative from IBM. (The potential conflict of interest of a member of the for-profit college prep industry serving on a higher education commission is noted.) Though part of the conservative George W. Bush administration, Spellings had endorsed a bipartisan vision for public policy and has represented the commission as non-ideological ("Margaret Spellings"). For a year, the commission worked to assess the state of the American college and university system, holding a series of public hearings and interviews with stakeholders in the higher education world. The report, officially named *A Test of Leadership: Charting the Future of U.S. Higher Education* but most often referred to simply as the "Spellings Commission report" or "Spellings report," expresses its fundamental question as "how best to improve our system of higher education to ensure that our graduates are well prepared to meet our future workforce needs and are able to participate fully in the changing economy" (33).

While announcing early on that the American university system has been the envy of the world for decades, the report shifts immediately to the threat posed by other higher education systems. "We may still have more than our share of the world's best universities," reads the report, "[b]ut a lot of other countries have followed our lead, and they are now educating more of their citizens to more advanced levels than we are… at a time when education is more important to our collective prosperity than ever" (*A Test x*). This competitive focus persists throughout the entire document. Again and again, the exigency for improving our colleges and universities is represented as a matter with keeping up with foreign powers. "Where once the United States led the world in educational attainment," the report warns, "recent data from the Organization for Economic Cooperation and Development indicate that our nation is now ranked 12th among major industrialized countries in higher education attainment" (*ix*). In contrasting supposed American educational stagnation with ascendant international competition, the Spellings Commission Report is part of the tradition of such reports contributing to a crisis narrative through such appeals.

The Spellings Commission called for reforms in five major areas: access, affordability, quality, accountability, and innovation. The area of most direct relevance to this project, and which has had the most immediate policy impact—and controversy— is accountability. In particular, the finding of direct relevance to the CLA is the call for standardized assessment measures in higher education, in terms of student outcomes and overall institutional quality. The report speaks of "a lack of clear, reliable information about the cost and quality of postsecondary institutions, along with a remarkable absence of accountability mechanisms to ensure that colleges succeed in educating students" (*vii*).

Throughout, the Spellings Commission report poses this lack of reliable information as the higher-order problem leading to the specific institutional and national problems within higher education. The result of these limitations in information, according to the report, "is that students, parents, and policymakers are often left scratching their heads over the answers to basic questions" (*vii*). The obvious solution to an information deficit is to find and deliver more information. However, the nature of that information—what is investigated and how—is a question of ideological and political weight. Here, the Spellings Commission is firmly on the side of standardization, calling for "outcomes-focused accountability systems designed to be accessible and useful for students, policymakers, and the public, as well as for internal management and institutional improvement" (24).

The report calls for several key elements that have become familiar elements of the recent assessment push: a focus on outcomes, a somewhat nebulous term that is invoked consistently in the assessment and accountability movement literature; the endorsement of value-added metrics, a controversial method of assessment that uses how individual and institutional scores change over time to assess educational quality (see Chapter 3, "History and Theory of the Collegiate Learning Assessment"); increasing access to, and standardization of, information available for students, parents, and the general public; and tying these reforms into accreditation. Throughout it all, the Spellings Commission report returns again and again to the need for standardization and standardized testing metrics. The report specifically suggested three standard assessment methods as models:

- the Collegiate Learning Assessment;
- the National Survey of Student Engagement and the Community College Survey of Student Engagement, a research effort of Indiana University designed to investigate how much time and effort students invest in learning at the collegiate level, and what the average requirements are for earning an American bachelor's or associate's degree;
- and The National Forum on College-Level Learning, a broad, multistate effort to understand college student learning, using such metrics as the CLA, the National Adult Literacy Survey, the two-year college learning assessment WorkKeys, and graduation admissions examinations such as the GRE, GMAT, and LSAT (*A Test* 22).

Although the report officially endorses no particular assessment, the CLA is mentioned three separate times as a good example of the kind of standardized assessment the Spellings Commission advocates. This highlighting of the CLA had a powerful impact on the visibility and viability of the CLA as a major assessment system.

The report does not merely advocate standardized tests as a method for achieving transparency and accountability, but also argues that there must be a system of incentives and penalties that makes this kind of assessment ubiquitous. "The federal government," reads the report, "should provide incentives for states, higher education associations, university systems, and institutions to develop interoperable outcomes-focused accountability systems designed to be accessible and useful for students, policymakers, and the public" (23). Perhaps keeping in mind the scattered and inconsistent policy

response to *A Nation at Risk*, the report here asks for federal intervention to ensure something resembling a coherent, unified strategy of assessment. The term "interoperable" is key. It suggests that states and institutions should not be made to conform to a particular assessment metric or mechanism, but rather to ensure that results from whatever particular assessment mechanism they adopt be easily compared to results from other mechanisms. This endorsement of local control and institutional diversity is common to American political rhetoric, where federalism and the right of local control are sacrosanct. As a practical matter, however, it is unclear whether there will really be a sufficient number of interoperable testing options to give states and institutions meaningful choices. The Spellings Commission also directed the regional accrediting agencies to go even further in pressuring colleges and universities to take part in rigorous assessment, instructing them to "make performance outcomes, including completion rates and student learning, the core of their assessment as a priority over inputs or processes" (24). This is the strongest message to the accrediting agencies yet delivered, calling on them not merely to make assessment of student learning a key part of their process, but their top priority. As in so many other parts of this history, the public good is invoked as the impetus behind major policy and procedural changes. "Accreditation," reads the report, "once primarily a private relationship between an agency and an institution, now has such important public policy implications that accreditors must continue and speed up their efforts towards transparency" (24).

As any document of this type would, particularly one commissioned by an extraordinarily controversial presidential administration like that of then-president George W. Bush, the report attracted considerable criticism. Most notable of all was

internal criticism. David Ward, the president of the American Council of Education, a consortium of accredited colleges and universities and various independent educational organizations, refused to sign the final report. At the commission meeting where votes were solicited, Ward was the only member to reject the report, although not the only one to express reservations. Saying that he was forced to "pour a little rain on this unanimous reaction to the report" (Lederman), Ward argued that the report's recommendations were too formulaic and specific to address the diversity of collegiate institutions or their unique problems. This response would come to be one of the loudest and most consistent complaints about the report. Additionally, he cited the tendency of the report to "to minimize the financial problems facing higher education but not of the industry's own making" (Lederman). Although the "no" vote of a single member had little impact on the commission, the lack of unanimous consensus was something of a speed bump. Additionally, Ward paved the way for more criticisms to come. The American Association of University Professors, the country's largest faculty union, cited Ward's refusal in its own response to the Spellings Commission. The report, argues the AAUP, "largely neglects the role of the faculty, has a narrow economic focus, and views higher education as a single system rather than in its institutional diversity" ("AAUP Statement").

Another commission member, Robert Zemsky, an education professor from the University of Pennsylvania, did formally sign the report. But years later, in a 2011 essay in the *Chronicle of Higher Education*, Zemsky expressed regret over having done so. In contrast with Ward's complaints, Zemsky argued that the commission's report was "so watered down… as to be unrecognizable" ("Unwitting Damage"). An initial

recommendation of the commission had been to develop a set of standard metrics that all colleges had to collect, but this effort was shot down by Congress, which asserted its right to regulate higher education. Congress's assertion of its authority to regulate colleges had the unfortunate consequence, in Zemsky's telling, of shifting the burden from the colleges and universities themselves to the accrediting agencies. That new scrutiny had the ironic effect of making colleges less likely to change; in order to placate the newly-defensive accrediting agencies, colleges became more formal and less transparent—directly undercutting the purpose of the commission. "Both irritated and alarmed, the accrediting agencies have done what bureaucracies under attack always do," writes Zemsky. "they have stiffened, making their rules and procedures more formulaic, their dealings with the institutions they are responsible for accrediting more formal and by-the-book… For a college or university now up for reaccreditation, the safe way forward is to treat the process as what it has become: an audit in which it is best to volunteer as little as possible" ("Unwitting Damage"). This criticism highlights a consistent feature of these kinds of top-down, sweeping reform efforts: their propensity, real or imagined, to result in unintended consequences.

The contradiction between those that see the Spellings commission report as too harsh and disruptive, and those who see it as too weak an ineffectual, is likely a result of the differing expectations and desires of the various observers. What is clear is that the consequences have already been wide-ranging, and are still being felt years after the publication of the report. These changes can be seen in the initiatives and policy decisions undertaken by the current presidential administration, that of Barack Obama.

The Obama Administration

Despite the fact that the Obama's election was explicitly positioned by his campaign as a break from the Bush administration, and the change in party control of the White House, the Obama administration's approach to higher education reform has not been as radically different from that of the previous administration as might be assumed. The major difference, as will be seen, comes in the degree of flexibility and local control on offer. Interestingly, the Republican Bush administration's approach to education was more top-down and national in its approach, reflected most obviously in the rigid, national standards of No Child Left Behind, while Obama's Race to the Top is more flexible and federalist in its approach. This difference turns traditional partisan assumptions on their head. Still, there has been remarkable continuity in education policy from the Bush administration to the Obama administration. That continuity, however, has occurred in a rapidly changing American economy.

Essential to understanding the higher education policy of the Obama administration is recognizing the financial crisis that immediately predated it and the steep recession that dominated its first several years. As has been discussed in countless books, articles, documentaries, and other media, the last year of the Bush administration witnessed an unprecedented crisis within the American finance industry, one that threatened the very foundations of our economy. A massive real estate bubble, driven by tax policies designed to encourage home ownership and by luxury development, raised the price of housing and along with it the value of mortgage-backed securities. Eager to sell more and more mortgages, given the profits raked in by selling speculative financial derivatives backed by the value of mortgages, banks and lenders pushed more and more

"subprime" mortgages onto low-income buyers who could not afford their payments. Eventually, the huge number of defaults caused a massive shock to the financial system, driving some of the largest investment banks, such as Bear Stearns, out of business. The ultimate result was a deep recession, one defined by massive job loss. According to the *Monthly Labor Review*, the US economy shed some 6.8 million jobs in 2008 and 2009, driving the unemployment rate to 11% and the average length of unemployment to 35 weeks (Kelter). In total, the financial crisis led to the worst American labor market since the Great Depression.

Workers with a college degree, as they long had, continued to enjoy both a wage premium and a significantly lower unemployment rate than the national average. In 2009, the first year of Obama's presidency, Americans holding a bachelors degree earned $1,025 a week and had an unemployment rate of 5.2%, compared to those with only a high school diploma, who made an average of $626 a week and had an unemployment rate of 9.7%, according to the Bureau of Labor Statistics ("Education Pays 2009"). This advantage, however, masked deep problems. To begin with, while the advantage in unemployment rate was impressive, the typical unemployment rate for college graduates has historically been below 4%, demonstrating that while the relative advantage over those without a college education was robust, in absolute terms the odds of a college graduate being unemployed had risen fairly sharply. What's more, these overall unemployment figures consider workers of all ages. A particular difficulty of this recent financial turmoil has been the unusual depth of the crisis for the youngest workers, recent high school and college graduates. In the post-financial crisis labor market, college graduates under the age of 25 reached a peak unemployment rate of above 9.5% in 2009

(Weissman). In other words, while recent college graduates maintained a lead over members of their own age cohort, their overall employment numbers were close to that of those with only a high school diploma across the age spectrum. Compounding matters was the explosion in student debt loads. The Project on Student Debt reports that, for the class of 2012 (who entered college in fall of 2008, at the beginning of the financial crisis), "[s]even in 10 college seniors… had student loan debt, with an average of $29,400 for those with loans" ("Student Debt and the Class of 2012" 1). In large measure, this student loan crisis was the product of rapidly increasing tuition costs. According to the College Board, in the decade spanning from 2002-2003 to 2012-2013, average tuition rates nationwide rose at a rate of 5.2% relative to inflation ("Average Rates of Growth"). In the early years of the Obama administration, then, college students were graduating with more debt than ever, into a punishing labor market that could not provide many of them with the kinds of jobs they expected to find.

Given this environment, there is little surprise that the Obama White House embraced the rhetoric of reform and accountability that was exemplified by the Spellings Commission report. In particular, the Obama administration has pushed hard for the collection and publication of more standardized information about colleges for parents and potential students. In his first administration, the bulk of the president's domestic policy was focused on the passage of the Patient Protection and Affordable Care Act (PPACA), popularly referred to as Obamacare, and on combating the deep economic malaise that afflicted the country. But in time, higher education reform would become one of the key aspects of his domestic policy. At a speech delivered at the University of Michigan at Ann Arbor in January of 2012, President Obama delivered one of the most

important statements of his education policy. In the speech, he called for a national effort by colleges and universities to curtail tuition increases, referring to this effort as a "Race to the Top" for college affordability. "Look, we can't just keep on subsidizing skyrocketing tuition," said the President. "And that means that others have to do their part. Colleges and universities need to do their part to keep costs down as well" ("Remarks by the President"). The notion that college tuitions are best kept low, of course, is a matter of little controversy. But Obama's speech went a step further, arguing that the federal government must tie access to federal funding to the ability of colleges and universities to keep tuition rates in check.

> from now on, I'm telling Congress we should steer federal campus-based aid to those colleges that keep tuition affordable, provide good value, serve their students well. We are putting colleges on notice – you can't keep – you can't assume that you'll just jack up tuition every single year. If you can't stop tuition from going up, then the funding you get from taxpayers each year will go down. We should push colleges to do better. We should hold them accountable if they don't. ("Remarks by the President")

This proposal marks a potentially massive change. By tying efforts to reduce tuition increases to access to federal funding, such as that used in financial aid and research grants, the White House proposal would create the first real enforcement mechanism for college affordability. As part of this enforcement mechanism, the president also called for a standardized college "report card," made available to the public, that reports both how affordable a given college is relative to peer institutions and how well its students are doing. In this, the program echoes the Obama administration's Race to the Top

program for K-12 schools, which similarly ties availability of federal funds to performance on college rankings. The relevance to standardized assessment is clear.

The broad outlines discussed in the speech were made explicit a year and a half later. In a fact sheet distributed to the media in August of 2013, the Obama White House laid out a multiple-point plan for college accountability. Among the points most important for assessment include

- Tie financial aid to college performance, starting with publishing new college ratings before the 2015 school year.
- Challenge states to fund public colleges based on performance….
- Give consumers clear, transparent information on college performance to help them make the decisions that work best for them. ("Fact Sheet" 2)

The proposal calls for legislation that will ensure that "taxpayer dollars will be steered toward high-performing colleges that provide the best value" (2). Which colleges are high-performing, in turn, will be based on the new series of ratings, which are to be calculated based on factors such as

- Access, such as percentage of students receiving Pell grants;
- Affordability, such as average tuition, scholarships, and loan debt; and
- Outcomes, such as graduation and transfer rates, graduate earnings, and advanced degrees of college graduates (3)

While the exact formula for these ratings remain to be seen, clearly, this proposal is the most direct and clear expression of external accountability yet put forth by a presidential administration. What's more, the proposal to tie federal aid to these ratings creates an

enforcement mechanism previously missing from past reform efforts. In its insistence on new, transparent assessments of college outcomes, the Obama proposal clearly interfaces well with the Spellings Commission report that came before it. Conspicuous in its absence from this document is an embrace of standardized assessments of student learning like the CLA. However, the fact sheet does endorse the possibility of "competency-based" approaches that reward students on performance rather than course hours. This might open the possibility for performance on a test like the CLA to be rewarded with college credits, as part of a broader competency-based aspect of college education. Where the Spellings commission advocated for a somewhat constrained definition of student success, the Obama administration's proposals seem to leave more room for flexibility.

Like the Bush administration before it, the Obama administration has been marked by near perpetual controversy. In contrast with his massively controversial overhaul of our nation's medical care system, the president's proposed reforms of higher education have attracted far less attention. Yet there has still been a great deal of discussion and debate about these proposals within the higher education community. Writing in *The Chronicle of Higher Education*, considered by many to be the most prominent news and opinion publication in American higher education, contributing editor Jeff Selingo praised the Obama proposal, comparing it favorably to the Obamacare health industry overhaul. "Right now, too many colleges are not getting the job done," writes Selingo, "whether it's not graduating enough of their students, especially those on Pell Grants, or putting too many of their students or their students' parents deep in debt in order to finance a degree with little payoff in the job market, today or five years from

now" ("President Sees an Obamacare Solution"). The Obama administration's proposals, writes Selingo, "are a start to rethinking what we want out of the vast federal investment in higher ed." A response of particular interest came from Margaret Spellings, whose commission generated the report that informed many of the Obama White House proposals. In an interview with *Inside Higher Ed,* Spellings was supportive of the general thrust of the proposal but questioned the practicality and efficacy of some of the details. "It's the right issue at the right time," Spellings said, "and I commend him for engaging on it" (Stratford). "Having said that, some of the proposals are unworkable and ill-conceived in the short run…. We need to start with a rich and credible data system before we leap into some sort of artificial ranking system that, frankly, would have all kinds of unintended consequences."

The Washington Post solicited the opinions of many prominent university presidents, obvious stakeholders on this issue. Their reactions were more mixed. Cornell University president David Skorton was generally positive, saying, "We need to give parents and students access to appropriate and robust metrics… so the overall idea is a good one" (Anderson). Similarly, Georgetown University president John J. DeGioia expressed support, saying, "Georgetown shares President Obama's commitment to increasing access and reducing the cost of higher education." However, Catholic University president John Garvey warned about federal intrusion into local control. "[O]ne of the questions we need to ask," says Garvey, "is how much deeper do we want the government to get into this business, if it means the government will also be calling the tune?" Meanwhile, Trinity Washington University president Patricia Macguire feared that the initiatives would in fact have the opposite of the intended effect. "Far from

helping us control costs," she argues, "this whole thing is just going to add a cost burden, add expenses to higher education." The most common reaction was exemplified by Morgan State University president David Wilson, who said, "The devil will be in the details, and the details about how this would work are not yet known." Sensibly, many of the college presidents, and commentators writ large, argued that the quality of the proposal was ultimately dependent on the quality and fairness of the metrics to be used in assessing college quality. "We must be very careful," said Wilson, "not to end up with a system of rating colleges and universities where institutions with plentiful resources are more advantaged than those without such resources. Certainly, if you accept a disproportionate number of students with stratospheric SAT scores, and if you have large endowments, such a rating system could become a cakewalk for those institutions." Part of the difficulty of effectively developing a set of fair and practically useful college rankings, then, is to establish egalitarian metrics for what are inherently inegalitarian institutions.

Conclusions

The Obama administration's efforts are still nascent, and the legislative and political battles ahead will likely be difficult. It remains to be seen what form the eventual ratings will take, or if they will survive political challenges at all. The initial proposals called for the creation of these rankings "before 2015," an ambitious goal that now appears unlikely to be met, with the adjustment of federal aid based on these rankings to take place by 2018. It is not clear whether that deadline will be met or if this system will ever be implemented at all. What is clear is that the message of accountability and a need for transparent assessment has fully taken hold of the conversation regarding higher

education. For good or for bad, the drumbeat of calls for national systems of accountability and assessment has grown to loud to ignore, for perhaps all but the most prestigious, economically independent institutions. From the early beginnings of federal review of higher education through to the present day, the case for regularized, interoperable systems of accountability has grown stronger and stronger. These calls are now a tacit part of the American collegiate landscape, and in casual conversation and academic scholarship alike, the debate is not so much whether universities will dramatically expand their assessment efforts, but what precise form that expansion will take. The constancy of these calls for reform, however valid those calls might have been, have given the assessment movement the seeming support of the weight of history.

The larger historical picture has also been made clear: national politicians engage with the question of higher education through a rhetoric of crisis and immediate exigency. While Reagan's *A Nation at Risk* took this crisis narrative to an extreme (and the contrary evidence compiled by Sandia Labs demonstrates the problems with this narrative most acutely), it is clear that the language of immediate exigency and dire problems is the default vocabulary of higher education reform efforts. National politicians simply find immediate causes and sources of current anxiety, usually tied to international competition from antagonist nations, and invoke them in calling for deep reforms of higher education. This crisis rhetoric does have the advantage of making the stakes clear, and in the best cases can rouse the legislative machine to provide more attention, and more funding, to our colleges and universities. But the downside of the crisis narrative is that it inevitably damages public perception of our institutions. Constantly claiming that our higher education system is in a state of crisis, even if the reasons and arguments change over

time, cannot help but create a weariness and unhappiness about that system in the public eye. That leaves our institutions vulnerable to political attack, and makes them incapable of defending themselves against aggressive reform efforts—reform efforts of the type the Obama administration is now pushing.

The enduring, essential question is whether any of these efforts will bear fruit. In order for all of this to work, the systems of assessment and accountability must be proven to assess student learning outcomes validly and reliably. At present, the assessment systems that are being utilized to fulfill the broad mandate for better understanding of college learning are largely ad hoc, lacking the kind of interoperability that the Spellings Commission calls for and that is necessary to have a truly reliable picture of student learning. Out of the available tests that could become national standards, the Collegiate Learning Assessment would seem to be in the best position to succeed, given its pedigree, its embrace by the national education reform movement, and the ambitions of its developers. If the CLA is to become a primary instrument in these accountability efforts, it will need to be demonstrated to accurately reflect real student learning, in a way that is acceptable to a large number of stakeholders, and in a manner that does not disadvantage students or institutions that lack the resources or prestige that some enjoy. In order to adjudicate these questions—to assess the assessment—I will explore the theoretical and empirical nature of the CLA and CLA+. That exploration is the subject of the next chapter.

CHAPTER 3 HISTORY AND THEORY OF THE COLLEGIATE LEARNING ASSESSMENT

As described in Chapter Two, the movement towards standardized (or interoperable) tests of collegiate learning has been building for some time. But the specific mechanism of the College Learning Assessment has its own history, one that must be placed in context with the beginnings of academic assessment and in comparison to similar and competing test instruments. The chapter that follows examines this history and context, and considers the theoretical, empirical, and practical realities of the CLA.

Early Precursors

Although the history of educational testing and assessment is too large to be adequately summarized in this space, it is important to reflect on some of the most important pioneers of this field, in order to place the CLA in an appropriate historical context. One of the earliest proponents of standardized assessments in higher education was Ralph W. Tyler, whose 1949 book *Basic Principles of Curriculum and Instruction* was the most prominent and influential such text of its time. Tyler, referred to by Stark and Lattuca as "the father of educational evaluation" (31), was not merely an early proponent of higher education assessment, but also of explicit learning goals and outcomes. In contemporary times, this emphasis on goals and outcomes may seem like an obvious facet of education, and yet in the traditional liberal arts curriculum, explicit

learning goals have not always been the norm. Tyler believed that in order to adequately assess learning outcomes, they had to be made specific. As he writes in *Basic Principles,* "if efforts for continued improvement [in educational outcomes] are to be made, it is very necessary to have some conception of the goals" (3). Tyler's book laid out a series of concerns and ideas for curriculum and assessment developers. Many of these today appear conventional now, but in the context of those early days, they represented a significant evolution in the study of testing and became part of the bedrock of educational theory. What is also clear in Tyler's text is a dynamic that has troubled educators and administrators ever since: the tendency for assessment needs to drive changes in curriculum needs, rather than the other way around. "These educational objectives," he writes, "become the criteria by which materials are selected, content is outlined, instructional procedures are developed and tests and examinations are prepared" (3).

Tyler identified several key aspects of effective educational assessment. Among them are

- Objectives of assessments must be realistic—that is, average students must have a reasonable expectation of being able to perform adequately on assessment tasks

- The assessment mechanism must provide students with a sense of accomplishment or emotional benefit (what Tyler calls "satisfactions") both on principle and because it was the only way to ensure student effort

- Assessments must be authentic, in that they match as closely as possible the actual skills and abilities that they are meant to assess, and in so doing

> "determin[e] the extent the educational objectives are actually being realized by the program of curriculum and instruction" (106)

- The mechanisms of assessment must be carefully designed to be coherent and iterative, so that the logical connections between tasks made them less frustrating for students and more useful for researchers, teachers, and administrators.

Again, these might seem like banal aspects of educational measurement and assessment. But Tyler, and early precursors like him, were only just developing norms and expectations for this nascent field.

A generation later, Hilda Taba was among the most influential researchers in education and curriculum working to expand and codify Tyler's earlier theories. Although Taba wrote her well-respected dissertation *Dynamics of Education: A Methodology of Progressive Educational Thought* in 1932, well before the publication of Tyler's book, her most influential work would be published decades later. Taba, a graduate of the Teacher's College of Columbia University and head of curriculum at the famous Dalton school in New York City, was among the first to articulate a need for more complex measurements to assess more complex learning goals. Her hallmark 1962 book, *Curriculum Development: Theory and Practice*, advocated strongly for a turn towards tests that could measure student abstract reasoning skills, rather than simple facts or figures. She argued that the assessments of her time created a "discrepancy between the scope of the objectives of curriculum and the scope of evaluation" (313). Taba was one of the first education scholars to articulate the idea of data obsolescence, the now-

ubiquitous notion that knowledge of facts, in and of itself, is of limited use to students. In a world with Google, this idea is now common, but Taba embraced it decades before the popularization of the internet. Taba pointed out that many facts can quickly change, but the process through which information is acquired and assimilated remains essential. Arthur Costa and Richard Loveall summarize Taba's requirements for the deeper mental processes that should be taught and assessed: "they must have scientific validity, they must be learnable at the age level at which they are offered, and must have utility in our current culture" ("Legacy of Hilda Taba" 58). As an example of the difference between facts and the abstract reasoning Taba saw as of greater importance, Costa and Loveall contrast the difference between knowing the current borders of Kenyan and Nigeria and knowledge like "national boundaries are created by many factors, including natural features, wars, and whim" (58). The former knowledge could easily go out of date; the latter will endure. In time, this thinking would be applied to the development of the CLA. Another of Taba's major influences lay in her contribution to the notion that individual academic skills could be disaggregated from broader learning and education. Her book referred to this philosophy as "componentality," in which various aspects of education could be divided into components in order to be studied, and that taken in aggregate, these components would represent an overall picture of the student's learning. Although this notion did not originate with Taba, her influential voice helped give credence to this view, which would grow to be the dominant position in education and assessment.

Though Tyler and Taba were only two of the many early practitioners of educational measurement and assessment, they were also two of the most influential, and two who best predicted the contours of future assessments. With their focus on practical

knowledge and their insistence on the ability of educators to measure student learning, both moved away from the traditional assumptions of the classical liberal arts education and towards the values that we can see in the assessment movement of today. Taba presaged the controversies of today in writing, "In a society in which change comes fast, individuals cannot depend on routinized behavior or tradition in making decisions whether on practical everyday or professional matters, moral values, or political issues. In such a society there is a natural concern that individuals be capable of intelligent and independent thought" (*Curriculum Development* 215). This attitude seems natural and pragmatic, but in time as discussed in Chapter 2, such concerns would lead to a perpetual crisis narrative about the university.

The Old Standards: The GRE and Similar Entrance Exams

The ideas and techniques developed by these pioneers would filter out into education and educational research in the 20th century, but these developments were largely centered on elementary and secondary education. In contrast, there was little organized development of assessments of higher education. Colleges and universities remained largely independent entities, free to dictate curricula and standards on their own. One of the few reasons college learning has been measured in the past has been for the purposes of determining which students are ready for graduate and professional education. In much the same way as the SAT is designed to tell colleges and universities which students are best prepared for post-secondary education, tests like the Graduate Record Examination (GRE), the Law School Admission Test (LSAT), the Graduate Management Admission Test (GMAT), and the Medical College Admissions Test (MCAT) are designed to assess which students are ready for various types of graduate education. The

most broad-ranging of these, and one taken by upwards of 700,000 students a year, is the GRE ("E-Update").

The GRE was originally developed in the late 1930s by a consortium of elite colleges, under the direction of the Carnegie Foundation, then as now a prominent philanthropic organization dedicated to developing policy and research about education. The tests were, in these early stages, content-based; that is, they assessed students on domain-specific knowledge in different disciplines. The test evolved fairly constantly through its first decade of existence, but by 1949, the GRE Aptitude test, which attempted to assess general cognitive skills and reasoning of college students, was born (*The Graduate Record Examinations Testing Program*). Although its name would change, and it would be tinkered with nearly constantly in its early years, the basic structure and function of the General GRE test had materialized: a test of reasoning and aptitude rather than content, divided into verbal and quantitative sections, used to assess how well prepared college students were for graduate study. By the beginning of the 1950s, another change would bring the GRE closer to the modern version: the Carnegie Foundation happily handed administration of the test over to the Educational Testing Service, the for-profit testing wing of the College Board, which by 1952 had adapted the test's scoring to fit the same 200-800 range, 500 average score system they had implemented on their SAT (Shavelson 29).

The GRE was joined in time by tests designed to assess student readiness for particular types of graduate education: the MCAT actually predates the GRE, having been first offered in 1928; the LSAT in 1948; the GMAT for business school applicants, in 1958. ETS itself would add additional subject-area specificity in the form of the GRE

Area tests (later Subject tests) in 1954. The exact subjects would vary over the years, with some being added and some discontinued, but in each case, the Subject tests were originally designed to offer students reasoning and evidence-evaluation tests within their specific field of interest. Currently, the GRE Subject tests offered by ETS are Biochemistry, Cell and Molecular Biology; Biology; Chemistry; Literature in English; Mathematics; Physics; and Psychology ("About the GRE Subject Tests"). Each of these field-specific tests have their strengths and weaknesses, but for obvious reasons, none functions as a practical test of general collegiate academic ability—they are subject-specific, and despite the breadth of options, there are many fields and majors unrepresented among them. This specificity and lack of breadth leaves the GRE General test as a kind of de facto leader in assessing college student ability, given the test's focus on domain-general reasoning skills and status as a general exam.

But despite its preeminence, the GRE has rarely been thought of as a candidate to assess programs and institutions. For one, there are consistent controversies and problems that have dogged the test for years. As with any test of this prominence and stakes, the GRE has been accused of being unfair, invalid, and insecure (Kaplan & Saccuzzo 303; Celis). Critics have long argued that the GRE General test does not actually predict student success in graduate education. A 1997 case study from the journal *American Psychologist*, for example, found that "the GRE was predicted to be of some use in predicting graduate grades but of limited or no use in predicting other aspects of performance" (Sternberg and Williams 630). In fact, the study found that only first-year grades were at all predictable from GRE results. Part of the difficulty with assessing the validity of a test like the GRE lies in the restricted range of grades found in graduate

education. Generally speaking, graduate grades are clustered at the top of the distribution. As ETS put it in a report defending the validity of the GRE, "graduate student grades are generally very high, and typically they show very little variation either within or across programs or institutions. The lack of variability of grades… creates a restriction of range that artificially limits the size of correlations that can be attained" ("What is the Value" 7). This lack of variability in grades points to a deeper problem with conceptualizing and measuring graduate student success, as that success is typically defined in harder-to-measure areas such as research and teaching quality. Another common complaint about the GRE is that it in fact measures general cognitive ability, and not educational aptitude or learning. (See, for example, Hunter and Hunter 1984.) This complaint would later also be levied against the CLA. (See "Validity" below.) Like the SAT and many other standardized tests, critics of the GRE have argued that the test is racially biased. A 1998 study from the *Journal of Blacks in Higher Education* found a large and persistent gap between black and white takers on the GRE, and argued that this gap could have major negative consequences, saying that "the evidence clearly shows that if admissions to graduate schools are made without regard to race and based largely on GRE scores, black students will be nearly eliminated from the graduate programs at the nation's highest-ranked institutions" ("Estimating the Effect a Ban" 82).

More important than these challenges to the validity and reliability of the GRE, however, is the fact that the GRE was never intended as an assessment of secondary education colleges and programs. The test has always been focused on evaluating students, rather than institutions. This problem is represented most acutely in the GRE's lack of control for ability effects—that is, the test does not have any way to demonstrate

student *growth*, only their final ability. Colleges, of course, differ significantly in the test scores, grades, and other markers of student success for their incoming students. The selectivity of the admissions process exists precisely to ensure that only the students with the most impressive resumes attend elite colleges. (Elementary and secondary education has similar problems, but these are typically the product of demographic issues like parental income and education level, and are less explicit and acute.) It's impossible for GRE scores alone to demonstrate how a student has grown during his or her time at a college, meaning that it is impossible to use such scores to assess the difference between an elite Ivy League institution and an open enrollment college; the differences in incoming ability are just too large. The CLA addresses this through its value-added model (see "The Slippery Measurement of Value Added" below). What's more, few college educators are likely to see the GRE as a valid test of higher learning. While there is a writing section and a few quantitative questions that ask students to supply their own answer, the large majority of GRE General Test questions are multiple choice. As Shavelson writes, "Faculty members [are] not entirely happy with multiple-choice tests…. They want[] to get at broader abilities, such as the ability to communicate, think analytically, and solve problems" (30). Clearly, if the higher education assessment mandate is to be fulfilled, a new measure of collegiate learning is required.

The Council for Aid to Education

The history of the Collegiate Learning Assessment is inextricably bound with that of the Council for Aid to Education (CAE), the New York City-based nonprofit that develops and administers the test. The CAE has a long and complex history, which is summarized on the CAE website under "History." The organization was founded in 1952

as the Council for Financial Aid to Education, under the directive of a set of corporate executives, chief among them Alfred Sloan, the CEO of General Motors. Sloan was already famous at that time for his leadership of GM, having pioneered many aspects of corporate governance and led GM into the modern era of automotive manufacturing. According to CAE, the purpose of this organization was to spur more charitable giving to colleges and universities, particularly among corporate entities, with a "goal was to increase the number of citizens who went to college" ("History"). For over thirty years, CAE participated in advertising and outreach campaigns to encourage charitable giving to institutions of higher learning. According to CAE, it was "first organization in the US to regularly provide national statistics on private giving to higher education" ("History"). In 1996, CAE became a subsidiary of the RAND Corporation, a well-known think-tank dedicated to applying empirical research and economic theory to social problems. In 1997, CAE contributed to the higher education crisis narrative by publishing a position paper titled "Breaking the Social Contract: The Fiscal Crisis in Higher Education." The paper argues that unsustainable growth in costs would make college unaffordable for many students, and would ultimately cause the higher education system to fail to meet growing demand. In 2005, CAE was spun off from RAND under its own leadership again. Since then, it has devoted most of its resources to the CLA initiative, although it also provides assessments for K-12 education, particularly in alignment with Common Core and state-based standards.

The Collegiate Learning Assessment

The CLA arose from a perceived lack of reliable tools to assess college learning. The most comprehensive history of the development of the CLA and CLA+ is

Shavelson's *Measuring College Learning Responsibly: Accountability in a New Era* (2010). Although the book was published too early to include information on the switch from the CLA to the CLA+, and was released before many of the schools that currently use the test adopted it, Shavelson's text is an essential document for understanding the philosophy, assessment mechanism, and history of the test. The book describes how the three most important developers of the CLA—Shavelson, Steven Klein, and Roger Benjamin—came together to create what would become the CLA. Shavelson and Klein are both psychologists by training, with research experience in developing assessments of student learning; Benjamin, a former dean and provost with a background in political economy (Shavelson 44). The three had long privately discussed the need for more transparency and accuracy in assessing the quality of education of various undergraduate colleges and universities. In his book, Shavelson reflects their frustration in writing that "information about learning was available…. But there was no way to benchmark how good was good enough" (44). With Benjamin's appointment to president of CAE in 1996, the group had the kind of institutional resources and clout to begin to turn those desires into a concrete reality. In the late 1990s and early 2000s, the three of them began to make the case for the assessment instrument that would eventually become the CLA. The most direct and strident of these calls was published in 2002 in the academic magazine *Peer Review*. Benjamin and his colleague Richard Hersh, another important progenitor of the CLA, wrote that "student outcomes assessment should be the central component of any effort to measure the quality of an institution or program" ("Measuring the Difference"). A year later, Benjamin and Marc Clum published another piece in *Peer Review*, titled "A New Field of Dreams: The Collegiate Learning Assessment Project," in which they

announced the CLA project and detailed some of its goals. In 2005, a team of five researchers led by Klein published "An Approach to Measuring Cognitive Outcomes Across Higher Education Institutions," which reported on the first real administrations of the CLA. In that year, data collection for the project began in earnest, and the CLA became a prominent part of the college assessment landscape.

The Performance Task

The central assessment mechanism of both the CLA and the CLA+ revision is the Performance Task. The Performance Task is a 60 minute, written-response task that presents students with a "real-world" scenario that requires them to make a decision and defend it using data, abstract reasoning, and argumentation. Every Performance Task prompt includes a description of the scenario, a summary of several points of view on the topic, and information presented in several different formats, such as tables, charts, and graphs. (Shavelson refers to this provided information as the "in-basket" (37).) Students role play the part of a key stakeholder in this decision, and must articulate not just why they made the decision they did, but what evidence and reasoning makes that decision best. The intent of the performance task is to demonstrate a student's ability to use various types of critical reasoning and argumentative skills in concert.

The website of the City University of New York, which has recently made adoption of the CLA one of its policy initiatives, summarizes the strengths of a quality Performance Task response as follows:

- Evaluates whether evidence is credible or unreliable
- Provides analysis and synthesis of the evidence
- Draws conclusions that follow from the provided evidence

- Is well-organized and logically developed, with each idea building upon the last
- Shows strong command of writing mechanics and vocabulary ("CLA Task Format")

This brief summary of the skills and abilities that should be demonstrated in a Performance Task response reflect broad implicit values and assumptions about the purpose of higher education. Central to this summary is the evaluation and use of evidence. As Shavelson notes, a key aspect of the Performance Task is knowing which evidence to use and how to use it. As he writes, "some of the information is relevant, some not; some is reliable, some not. Part of the problem is for the students to decide what information to use and what to ignore" (37). This focus on weighing and incorporating evidence intelligently is part of the effort to make the CLA a valid test across different majors and types of institutions. Rather than utilizing knowledge they already know, which would necessarily be subject to discipline-specific education and the idiosyncrasies of particular institutions, the CLA presents information of variable quality and relevance for the student to choose from and utilize as needed.

Human raters have always graded the Performance Task, although the developers previously assumed that this task would have been handed off to computers by 2010 (Shavelson). Raters score utilizing a rubric, which is divided into three components: Analysis and Problem Solving, Writing Effectiveness, and Writing Mechanics. These sections are defined in the following ways:

Analysis and Problem Solving. Making a logical decision or conclusion (or taking a position) and supporting it by utilizing appropriate information (facts, ideas, computed values, or salient features) from the Document Library

Writing Effectiveness. Constructing organized and logically cohesive arguments. Strengthening the writer's position by providing elaboration on facts or ideas (e.g., explaining how evidence bears on the problem, providing examples, and emphasizing especially convincing evidence

Writing Mechanics. Demonstrating facility with the conventions of standard written English (agreement, tense, capitalization, punctuation, and spelling) and control of the English language, including syntax (sentence structure) and diction (word choice and usage) ("Rubric")

The Analytic Writing Section

The second major section of the original CLA was the Analytic Writing section. This section was made up of two essay responses, one which built its own argument based on a typical short-writing prompt, and another which asked students to critique an argument. The former section was allotted 30 minutes, the latter 45 minutes. These essays were judged on their presentation, development, and persuasiveness, which corresponded to the clarity and concision of the argument, the effectiveness and logic of the writing structure, and the presentation and analysis of evidence, respectively (Shavelson 53). In many ways, the Analytic Writing was reminiscent of similar standardized timed essay tests such as those found in the SAT and GRE. Shavelson stresses that this test "depend[ed] on both writing and critical thinking as integrated rather than separate skills" (52). One notable aspect of the Analytic Writing is that it was scored by computer. Little

or no information is available about which automated essay scoring system was employed in the evaluation of these essays.

From CLA to CLA+

After implementing the CLA at hundreds of universities from 2007-2012, the CAE implemented the first major revision of the examination in 2013, dramatically changing the form of the test and with it, its name, adopting the new moniker Collegiate Learning Assessment+. The Performance Task has remained essentially unchanged. However, the developers dropped the Analytical Writing task entirely, leaving the rubric items concerning student writing within the Performance Task as the only test of student writing in the assessment. Dropping the Analytical Writing section also means that no portion of the test is now scored by computer. In the place of the Analytical Writing now stands the Selected Response section, wherein students answer questions by choosing from a list of prewritten responses. The CAE summarizes the Selected Response section as follows:

> In the Selected-Response section, students respond to 25 questions: 10 assess scientific and quantitative reasoning; 10 assess critical reading and evaluation; and 5 assess the students' ability to critique an argument. Students complete this section within 30 minutes. Much like the Performance Task, each set of questions requires that students draw information from accompanying documents. ("CLA+ Sample Tasks")

An additional change involves switching from the sample-and-infer method described above to a census-style approach where all students are tested on all items.

The changes from the CLA to the CLA+ are interesting and invite scrutiny. In particular, they are worth considering given the ways in which they deviate from the prior attitudes of those involved in the development of the CLA. First, with the demise of the Analytical Writing section, the CLA+ is 100% human scored. This is a marked change from the previous assumptions of CAE, which previously operated under the assumption that the test would, at some point, be scored entirely by computers. As Shavelson writes in *Measuring College Learning Responsibly*, "Currently, human judges score students' [Performance Task] response online, but by 2010, the expectation is that responses will be scored by computer" (38). But the Performance Task is still scored by trained human raters, and while the Selected Response section is presumably scored automatically, there is a clear difference between the kind of natural language processing and computerized analysis necessitated by automated scoring of written responses like that in the Performance Task and Analytical Writing and the rote checking of multiple-choice answers like that in the Selected Response section. The failure to adopt universal computer scoring as planned may simply be a matter of available technology failing to satisfy expectations. While automated scoring systems for student essays have continued to be developed, so too have criticisms and critiques of such systems. It's also important to say that while many argue for the value and use of automated essay rating software generally, what the rubric of the CLA Performance Task requires is the ability to judge complex constructs such as quantitative reasoning, argumentative and stylistic clarity, and rhetorical force. Even many proponents of automated essay scoring would be skeptical of the ability of extant systems to perform this kind of judgment. As Mark D. Shermis put it in an interview with *US News and World Report*, automated essay scoring

"can't tell you if you've made a good argument, or if you've made a good conclusion" (Haynie).

The adoption of the Selected Response portion of the test is telling in and of itself, given the degree to which it amounts to walking back prior commitments of the CAE. Prior to the development of the CLA+, CAE personnel made statements questioning the effectiveness and validity of multiple-choice testing. As Shavelson writes in *Measuring College Learning Responsibly*—only one of many moments in which he criticizes multiple-choice testing—"There are no multiple-choice items in the assessment; indeed, life does not present itself as a set of alternatives with only one correct course of action" (49). The choice of the name "Selected Response" may itself be an attempt to distinguish the task from conventional multiple-choice testing, even though there is very little to distinguish the Selected Response task in actual application. CAE documentation perhaps reveals a defensiveness about this change, as a pamphlet about the CLA+ argues that "[The Selected Response items] are far from the typical recall and recognition multiple-choice items seen in many other standardized assessments" ("Reliability and Validity" 3). It is unclear from CAE documentation why this major change occurred. CAE's website does mention that the CLA+ "enhance[s] the richness of the results we provide to institutions (and to students)" by "introducing additional subscores (scientific and quantitative reasoning, critical reading and evaluation, and the ability to critique an argument) to complement the subscores we've provided all along" ("Comparing CLA to CLA+). Speaking speculatively, it may be that institutions requested these types of scores be included in the CLA assessment, and CAE thought it necessary to introduce conventional multiple-choice testing in order to generate them. In any event, this change

likely demonstrates the degree to which previously essential commitments on the part of CAE have become flexible when faced with institutional and market pressures.

A final change, and again one which represents a significant walking back of prior CAE commitments, is the abandonment of sampling as a responsible method of evaluating student learning. In early literature about the test, developers sensibly argued that student populations could be responsibly sampled and, using straightforward processes of inferential statistics, represented statistically in a valid and useful way. As Shavelson writes in *Measuring College Learning Responsibly*, "The CLA also uses sampling technology to move away from testing all students on all tasks… The focus then [in earlier higher learning assessments] was on individual student development; CLA focuses on program development" (35). Later, he reiterates this point, saying that the CLA "focuses on campuses or on programs within a campus—not on producing individual student scores" (47). In this use of random sampling and inferential statistics to draw responsible conclusions about larger student populations, the CLA was both progressive and practical. The notion of "standardized test overload" has been a consistent controversy of the broad American education reform movement. (See, for example, "Headline News: 'Our Kids are Tested to Death," by the National Center for Fair and Open Testing.) What makes these concerns especially troubling is that, with the use of responsible sampling and inferential statistics, testing all students is unnecessary. The CLA's embrace of these techniques helped to reduce the testing burden on students while still giving institutions strong information about student learning. What's more, the sampling mechanisms of the CLA made the task of recruiting students to take the test—

and getting them to invest serious effort in it (see "Validity and Reliability" below)—easier on institutions.

But with the CLA+, this commitment has largely been abandoned. Indeed, the CAE is now using the census-style approach as a marketing tool. In a document detailing the differences between the CLA and CLA+, the CAE states that "[p]erhaps the greatest enhancement—the 'plus,' if you will—is the move to a version of the assessment in which all students take all components of the CLA+" ("Comparing CLA to CLA+" 1). Why the change? In part, this change could reflect market forces—some institutions may have indicated that they would rather use a census approach than a sampling approach. As with the adoption of the additional subscores detailed above, the move away from limited sampling is likely a change that was undertaken with an eye to institutional desire. Though this approach is more expensive, it is better in keeping with the broad movement for more comprehensive testing, such as is typical of state tests in K-12 education. Census-style testing also satisfies the spirit of the national collegiate assessment push of which the CLA is a part (see Chapter 2). The other major element of this switch reflects the CAE's ambitions that CLA+ test scores become a nationally-recognized marker of a student's performance—a kind of "SAT for college" that employers and graduate schools could use in weighing a job or admissions candidate. The CAE website says,

> Now with CLA+, new student-level metrics provide guidance to students and data to faculty and administrators for making decisions about grading, scholarships, admission, or placement. Institutions can use CLA+ for additional admissions information for college applicants, to evaluate the strengths and weaknesses of entering students. Results for graduating seniors may be used as an independent

> corroboration of the rapid growth of competency-based approaches among colleges. Graduating seniors use their results—issued in the form of verified digital badges—to provide potential employers with evidence of their work readiness skills. ("CLA+ Overview")

The CAE has made little secret of their ambitions: to develop and implement a test that becomes seen as one of the major benchmarks of early life success, in much the same way that SAT scores have done for teenagers for decades. This effort will be enormous, and it remains to be seen if students, schools, and employers will ever invest in the test sufficiently to make this kind of metric as ubiquitous as the CAE hopes. But with the positive mentions of the CLA in the Spellings commission report, and the growing chorus calling for higher education accountability, they have a head start.

Validity

One of the most important concepts for evaluating any measure of educational performance is validity. Validity, in the social sciences, refers to whether a given measurement accurately measures what it intends to measure. In his book *Practical Language Testing* (2010), Glenn Fulcher writes that "the key validity question has always been: does my test measure what I think it does?" (19). Although this question is straightforward, its answers are multiple and complex, particularly in contemporary research. For decades, the simple notion of validity described above, now known as "construct validity," predominated. But in recent years, the notion of validity has been extended and complicated. For example, predictive validity concerns whether performance on one test can accurately another variable, such as a student's SAT scores predicting first-year GPA. Criterion validity concerns whether a test or variable

accurately predicts a future competency or skill, such as if the results of a driving test accurately predicts whether a driver will be in a car accident. There are many more types of validity that have been identified and explored by researchers, such as convergent validity, which demonstrates how traits theoretically presumed to be related are actually related, and discriminant validity, which demonstrates how traits theoretically presumed to be unrelated are actually unrelated. These various, sometimes contrasting types of validity demonstrate why evaluating a test can be a formidable task.

The extant literature on the validity of the CLA is limited, with much of it emerging from CAE itself. A pamphlet provided by CAE called "Reliability and Validity of CLA+" argues that the test has construct validity thanks to self-reported survey results from students who had taken the test. These students were asked how well the test measured writing, reading comprehension, mathematics, and critical thinking and problem solving. In writing, reading comprehension, and critical thinking and problem solving, a clear majority of students felt that the test measured their ability at least moderately. However, fully 55% of students felt that the test did not measure mathematics well at all, perhaps reflecting the fact that the CLA+ does not have a section of direct mathematics questions typical to standardized tests. Overall, the pamphlet argues that these responses demonstrate construct validity for the test and that "it appears that we are measuring what we purport to measure on the CLA+ tasks" (5). This survey is encouraging, but it is fair to ask whether students who have no background in test development or research methods can adequately assess whether a test they took is accurately measuring what it intends to measure.

One of the most important tests of the CLA's validity is found in a larger study that considers several major tests of college learning: the CLA, ACT's Collegiate Assessment of Academic Proficiency (CAAP), and ETS's Measure of Academic Proficiency and Progress (MAPP). This 2009 study was undertaken under the auspices of the Fund for the Improvement of Postsecondary Education (FIPSE), a subsidiary of the Department of Education that provides funding for research in college education. The study was in fact authored by employees of CAE, ACT, and ETS. For this reason, it cannot be considered truly independent research, but the federal oversight and combined expertise from these different organizations enhance the credibility of this research. 1,100 students from 13 colleges took part in the study. When viewed on the school level, which lowers the variability in comparison to looking at the individual level, the correlations between all tests were generally high, ranging from .67 to .98 (Klein et. al. 2014 24). This indicates that the tests are measuring similar constructs, lending evidence to the concurrent validity of these tests. It is worth pointing out, however, that while this research indicates that all of these tests may be measuring similar qualities, that does not necessarily mean that they measure what the purport to measure, or that their measurements are free from systemic biases or lurking variables. It's also interesting to consider the high correlations between these tests in light of CAE's desires to differentiate their own test. While the organization has obvious interest in demonstrating that their test instrument is different from its competitors, they still take advantage of their test's similarity to these competitors to prove the validity of the CLA+.

An important and difficult question for evaluating tests concerns student motivation. A basic assumption of educational and cognitive testing is that students are

attempting to do their best work; if all students are not sincerely trying to do their best, they introduce construct-irrelevant variance and degrade the validity of the assessment. This issue of motivation is a particularly acute problem for value-added metrics test the CLA, as students who apply greater effort to the test as freshmen than they do as seniors would artificially reduce the amount of demonstrated learning. At present, the CLA is a low stakes test for students. Unlike with tests like the SAT and GRE, which have direct relevance to admission into college and graduate school, there is currently no appreciable gain to be had for individual students from taking the CLA. Frequently, CLA schools have to provide incentives for students to take the test at all, which typically involve small discounts on graduation-related fees or similar. The question of student motivation is therefore of clear importance for assessing the test's validity. The developers of the test apparently agree, as in their pamphlet "Reliability and Validity of CLA+," they write "low student motivation and effort are threats to the validity of test score interpretations" ("Reliability and Validity of CLA+"). Measuring motivation, however, is empirically difficult. One attempt was made at Central Connecticut State University, a CLA school. Dr. Brandon Hosch attempted to measure student motivation by examining how much of the 60-minute maximum test takers used, and comparing that time usage to SAT-normed scores. While Hosch acknowledges that there are some problems with using time-on-task to measure motivation, he finds that "when controlling for academic inputs by comparing actual CLA scores to expected CLA scores, a similar pattern emerges; students who spent more time on the test outperformed their expected score" (7).

Hosch also gave students a survey to report their level of motivation. While self-reported data must be taken with a grain of salt, Hosch found that only 34% of freshman

agreed or strongly agreed that they were highly motivated on the CLA (8). Seniors, on the other hand, agreed or strongly agreed that they were highly motivated 70% of the time. In their own surveying, CAE found that 94% of students rated their own motivation as moderate or above, although only 15.2% said that they made their best effort ("Reliability and Validity of CLA+" 4). Hosch suggests that his research indicates that "CLA (and likely other instruments) may exhibit sensitivity to recruitment practices and testing conditions… the extent to which these difference may affect scores presents opportunities to misinterpret test results as well as possibilities that institutions may have incentives to focus efforts and resources on optimizing testing conditions for a small few rather than improving learning for the many" (9).

Student motivation was also at issue in a major paper authored by researchers from ETS. In this 2013 study, Ou Lydia Liu, Brent Bridgeman, and Rachel Adler studied the impact of student motivation on ETS's Proficiency Profile, itself a test of collegiate learning and a competitor to the CLA+. They tested motivation by dividing test takers into two groups. In the experimental group, students were told that their scores would be added to a permanent academic file and noted by faculty and administrators. In the second group, no such information was delivered. The study found that "students in the [experimental] group performed significantly and consistently better than those in the control group at all three institutions and the largest difference was .68 SD" (Oiu, Bridgeman, Adler 356). That effect size is quite large, indicating that student motivation is a major aspect of such performance metrics, and a major potential confound. It is true that the Proficiency Profile is a different testing instrument than the CLA, although Oiu, Bridgeman, and Adler suggest that this phenomenon could be expected in any test of

college learning that is considered low stakes (359). The results of this research were important enough that Benjamin, in an interview with *Inside Higher Ed*, said that the research "raises significant questions" and that the results are "worth investigating and [CAE] will do so" (Jaschik). Clearly, the impact of student motivation on CLA results will have to be monitored in the future.

Reliability

Reliability refers to a test's consistency: does the test measure different people in different contexts at different times in the same way? A test or metric is considered reliable if, given consistency in certain testing conditions, the results of the test are also consistent. This means, for example, that students in different locales or time periods but of equal ability in the tested construct will receive similar scores on the test. An unreliable test can't be used fairly; if the test does not evaluate different people consistently, then it could result in outcomes that are not commensurate with ability.

For testing instruments like the CLA, one of the primary means of establishing reliability is with test-retest reliability. The assumption behind standardized assessments is that they reflect particular abilities and skills of the students being tested, and that these abilities and skills extend beyond the particular test questions and instruments. That is, while we should expect some variation from test administration to test administration for a particular test taker, a reliable instrument should produce fairly consistent results for a given scorer, absent student learning. A test taker should not score 1.5 standard deviations above the median score one week and 1.5 standard deviations below the median the next. Such a result would severely undermine our faith in the test's ability to fairly reflect that student's ability. In order to assess test-retest reliability, the CAE ran a

pilot study utilizing the original CLA assessment. The sample size of this pilot study is unknown. On a per-student basis, CAE admits, the test has only moderate test-retest reliability, in the .45 range ("Reliability and Validity of CLA+" 3). They attribute this low reliability to the paucity of information, as "at the individual student level, the CLA was only a single PT or Analytic Writing Task" (3). This is a strange defense; while it's true that a longer test with more items will frequently result in higher test-retest reliability, the pilot study utilized the real test instruments of the CLA. Future students will take the same Performance Task and given a score based in part on that instrument, and it's reasonable to ask whether repeated administrations of that instrument will result in consistent scores. The test fared much better on test-retest reliability when looked at from the institutional level. That is, did an institution's average or median CLA scores from one administration predict the average or median scores from the following administration? Here, the test performed much better, with a reliability of .80. This measurement suggests that there is strong but imperfect consistency in a school's average performance on the test, with the remaining variability likely reflective of differences in student ability and nuisance variables.

Another important component of test reliability is internal reliability, measured with Cronbach's alpha. Internal reliability refers to whether a test is a consistent measure of a given construct throughout its section. For example, a student who is excellent at math generally should be expected to perform well on math items throughout the test, and not just on one half of a test. Performance on different items that test the same constructs is expected to vary somewhat, and perfect consistency across items would suggest that these items are redundant. But generally, test takers should be expected to perform

consistently on items that test the same constructs. This consistency is typically measured using Cronbach's alpha, a coefficient that ranges from 0 to 1, with 0 representing no consistency in performance on test items and 1 representing perfect consistency in performance on test items. Generally, test developers attempt to achieve Cronbach's alpha scores of between .75-.95, which indicates high consistency in performance on items but not perfect consistency. In CAE's pilot study, they found "reliability was between .67 and.75 across the four [Performance Tasks.] Reliability for the [Selected Response Items] (α=.80 and .78) is higher than the PTs" ("Reliability and Validity" 3). These reliability coefficients are both fairly low in context with other tests, but still within the conventionally-defined acceptable range. It is not surprising that the multiple-choice items are more internally consistent than the Performance Task sections, given how much more variability there is in potential responses to the Performance Task prompts.

Criterion Sampling and Psychometric Assessment

One of the most consistently identified and important theoretical stances in the CLA literature lies in the concept of criterion sampling, or the belief that intellectual and academic abilities work together in concert and cannot be usefully separated through testing. The developers of the CLA explicitly and repeatedly define the CLA's criterion sampling in opposition to the traditional psychometric school of assessing learning, which assumes that such separation is possible. "This [criterion sampling] approach assumes that complex tasks cannot be divided into components and then summed," writes Shavelson, "[t]hat is, it assumes that the whole is greater than the sum of the parts and that complex tasks require the integration of abilities that cannot be captured when

divided into and measured as individual components" (48). The developers argue, therefore, that the various cognitive and academic abilities that the developers identify as keys to success on the CLA cannot be thought of as discrete skills to be understood separately. "To pull the tasks apart," write Klein et al., "and index critical thinking, analytic reasoning, and communication separately would be impossible with the complexity and wholeness of these tasks" ("CLA Facts and Fantasies 421). In other words, the CLA is a complex assessment for a complex educational world, and its parts are interconnected in such a way that they cannot be disaggregated into separate skills.

Discussion of criterion sampling, as an alternative to psychometric testing, can easily be confused by the divide between criterion referencing and norm referencing. Criterion referencing refers to tests and assessments in which test subjects are not placed on a scale relative to each other but rather are found to satisfy some criteria or another. A driving test is a classic example; test takers do not receive a score but are rather found to be either competent to drive, according to specific criteria, or not. Norm referencing, in contrast, involves assigning test subjects a score that can be compared to those of other test takers, allowing test developers to compare performance in terms of means, medians, standard deviations, and the like. While there is clearly some overlap in these concepts, it is important to be clear that the discussion in this section of this dissertation focuses on a theoretical conflict concerning whether intellectual and academic abilities can be meaningfully isolated and scored independently, rather than the differences between testing to meet a particular criteria and testing to locate a test subject on a scale.

The CLA's criterion sampling approach marks a major departure from most standardized tests, which are largely descended from the psychometric philosophy. The

psychometric assumption that intellectual and cognitive abilities can be subdivided into discrete parts stretches back to the formative days of intelligence testing. Fulcher locates the rise of psychometric theory and assumptions to the beginnings of the 20th century, and in particular identifies World War I as a major impetus in the need for tests of intellectual ability (16-17; 33). In this telling, changes to the nature of warfare, including the increasing dependence on complex machines and the integration of reconnaissance into combat, caused military officials to place a premium on intelligent personnel. This in turn required the creation of effective tests to determine which soldiers and officers were more intelligent, and led to an "explosion of testing theory and practice during the Great War" (Fulcher 33). These tests were generally psychometric in their approach, utilizing what is still sometimes referred to as "trait theory," which presumes that cognitive and communicative skills can be both effectively defined by researchers and test developers and separated from broader contexts. This presumption was largely tacit, without much theoretical justification. As Fulcher writes, "For early testers there was therefore no question that using tests was in principle no different from using scientific instruments to investigate natural phenomena" (33). Distinct cognitive skills could therefore be separately investigated as easily as distinct organs in the human body. This presumption underlies a great deal of the theoretical and empirical work in assessment over the history of the discipline. For example, the psychometric tendency can be seen clearly in Taba's *Curriculum Development*. The CLA's criterion sampling can thus be seen as a major departure from typical academic testing.

Of course, the practical question for test developers isn't merely whether to try to assess skills separately or in concert, but how to test effectively. Traditional psychometric

attempts to subdivide intellectual abilities stem in part from the perceived need to focus on specific constructs in order to make them easier to define and test. This identification of skills to be tested is typically referred to as "construct definition," and there is a vast theoretical literature that explores it. Construct definition is considered one of the most important aspects of test development. As Fulcher writes, "the definition of the construct is something that has to be undertaken carefully if it is to be assessed, as we need to know what it is we have to ask a learner to do, so that we can observe it , and decide whether (and to what extent) this abstract ability is present" (97). Given this importance and this need for care, it's easy to understand the tendency to test for smaller, more narrowly-defined constructs. As Taba writes, "the more important and complex the objectives, the less likely it is that there are systematic and dependable devices for measuring their achievement" (314). In other words, while it may be more natural and useful to evaluate student academic abilities in concert, as the CLA attempts to do, doing so also increases the challenge of testing well. The validity and reliability of the CLA's Performance Task is described above. The question is whether the manner in which the test assesses is actually consistent with the criterion sampling philosophy.

The approach taken with the CLA is fairly typical of written assessments: trained raters are given a detailed rubric that subdivides each Performance Task response into various components (see "The Performance Task" above). The obvious question is how the use of a subdivided rubric maintains the spirit of the criterion sampling approach detailed by Shavelson and other developers of the CLA. Since the CAE itself subdivides the Performance Task into Analysis and Problem Solving, Writing Effectiveness, and Writing Mechanics, and these sections further identify traits like logic, utilizing

information, elaborating on facts, and syntactic control ("CLA+ Scoring Rubric), it seems that the test developers do recognize that academic skills can be subdivided and assessed separately. I don't doubt that the test developers believe in their own theoretical rationale for attempting to assess these skills together rather than separately. But for practical reasons of best practices in testing, the identification of subskills appears to be necessary. In order to make tests of written responses reliable, raters must be given detailed information about how to assess those responses. This seems to inevitably require the identification of discrete skills in a way that cuts against the criterion sampling approach. These difficulties do not make the criterion sampling approach invalid, or mean that the CAE is wrong to attempt to assess skills together. But it points to the ways in which assessment theory and its various requirements dictate test development, sometimes against the preferences of the developers.

The CLA and the SAT: Is Another Test Necessary?

One of the consistent criticisms of the CLA in its history has been its high correlation with SAT results. Part of the difficulty in measuring educational quality lies in the profound impact of differences in student populations. If one teacher teaches a class with much higher initial ability, another a class of much lower initial ability, and these teachers are compared simply via average scores, the lower will likely appear to be worse even if he or she did a better job teaching. This discrepancy is an especially acute empirical problem in the context of American colleges and universities, which are explicitly and intentionally unequal in the incoming ability of their students. Elite colleges and universities invest enormous resources in finding and attracting the brightest, best-prepared students. Open access universities, in contrast, will take essentially any

students that apply. A difference in prerequisite ability is not just a possibility in higher education assessment but an inevitability. It's for this reason, in part, that the CLA is a value-added instrument; by comparing freshman to senior average scores, student growth can perhaps be measured, rather than just overall student ability. Another method to control for differences in student ability is with the use of SAT norming, which has been utilized the CLA and others, such as Arum and Roksa in their *Academically Adrift*. In this process, institutional-average CLA scores are regressed along SAT scores. Since these SAT scores are earned before a student even steps foot in college, they are a reasonable way to assess incoming ability without the influence of college learning. A scatterplot of both freshman and senior administrations of the CLA regressed on SAT scores is below as Figure 1.

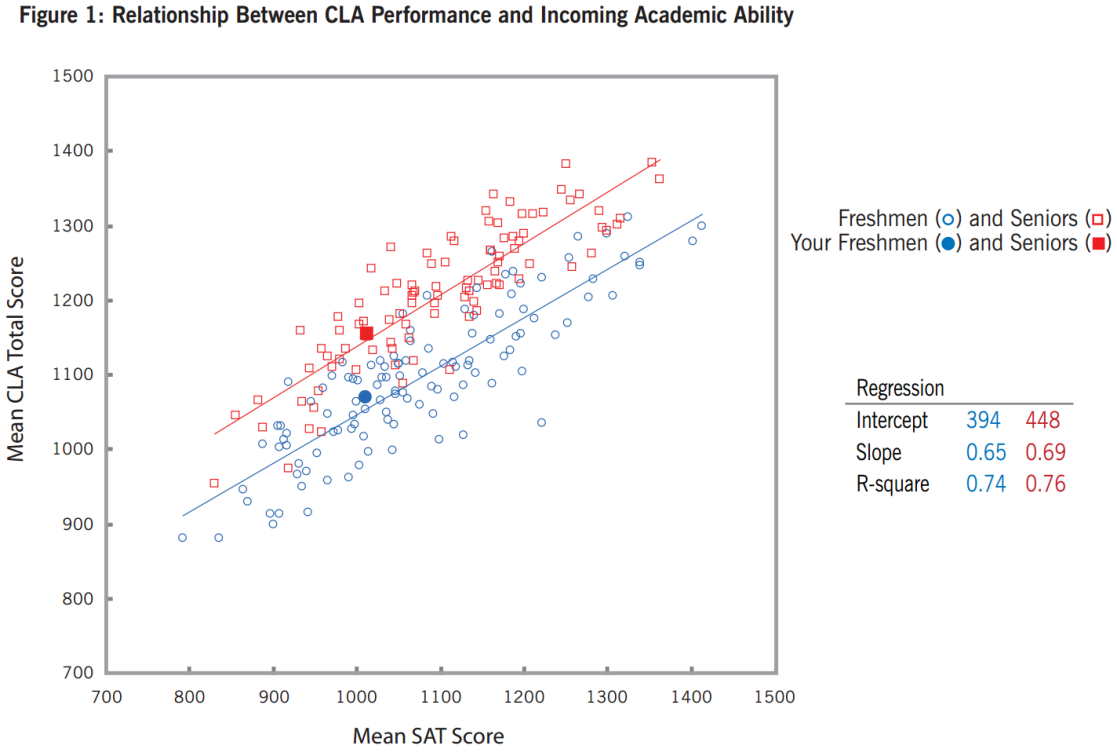

Figure 1 Relationship Between CLA Performance and Incoming Academic Ability

In this scatterplot, average scores for an institution's freshman class are represented as blue circles, and average scores for an institutions senior class are represented as red squares. The blue and red regression lines demonstrate the relationship between incoming SAT scores and CLA scores. It's important to understand that these are institutional averages; if individual student CLA scores were regressed on individual student SAT scores, we could expect far more variation in the scatter plot and a weaker relationship. As can be seen, in both freshman and senior administrations, SAT scores are strongly predictive of CLA scores, with an R-square value of .74 for the freshmen and .76 for seniors. This means that about 75% of the variation in CLA scores can be explained by SAT scores, and thus by incoming student ability, in this data set. In other words, we can predict 75% of an institution's average CLA score simply by looking at the SAT scores of its students.

This correlation, and others like it at the student level, have been the source of consistent criticism of the CLA: since SAT scores are so highly predictive of student performance, how effective is the test as a test of *college* learning, really? And why should time and resources be devoted to testing if SAT scores are so highly predictive of CLA scores? At some institutions, the relationship between the SAT and CLA is even stronger than in the above figure. For example, Steedle (2010) found that the correlation was as high as .93 in his research ("Incentives, Motivation, and Performance" 19). As Trudy Banta and Gary Pyke (2012) write, "given the strength of these relationships, it would appear that what is being measured is the entering abilities and prior learning experiences of students at an institution" (28). The developers of the CLA have disputed

this argument. In their 2007 article "CLA Facts and Fantasies," Klein, Benjamin, Shavelson, and Bolus attempt to refute this line of thinking:

> high correlations do not imply two tests are measuring the same thing—i.e., the same thinking or reasoning, or "cognitive processing." Consider the following. If we were going to teach to the CLA, our CLA preparation course would look a lot different from that used by Princeton Review to train students to take the SAT. That is, if college instructors trained students to work through CLA tasks, they would be teaching the kind of critical thinking, analytic reasoning, and communication skills their colleges' mission statements say they are teaching. Or put another way, even if the CLA correlates highly with the SAT, we would not take a student's SAT score as a proxy for his or her grade or performance in, say, an American history course—the SAT doesn't say anything about the student's history knowledge and reasoning just as the SAT doesn't say anything about a student's reasoning on the CLA. (430)

This rebuttal is reasonably sound thinking, but not quite persuasive. It is certainly true that tests of different cognitive or academic abilities can be consistently correlated without those tests measuring the same things. Although there are exceptions, generally, students who are strong in some academic areas relative to peers are strong in other academic areas. Scores on the SAT Verbal section and the SAT Math section are highly correlated for individual students, for example, with internal ETS research indicating a Pearson correlation of .71, a moderately high correlation (Dorans 18). It's certainly possible, therefore, for the SAT and CLA to test distinct variables and constructs without being redundant. The question, however, is whether these high correlations confound the

ability of the CLA to truly measure collegiate learning. While we could assume that the SAT and CLA are testing different things, knowing that a test taken in high school is so highly predictive of CLA results undermines our ability to trust that the results of the test are a matter of college learning and not incoming ability. We are potentially left with a test that is unable to distinguish between the effectiveness of college teaching and the exclusivity of that college's selection process. The most important mechanism that CAE utilizes to combat this problem is in the measurement of value added.

The Slippery Measurement of Value Added

Value-added metrics in educational testing attempt to compensate for differences in prerequisite ability by comparing pre- and post-test results to show how students have grown from one test administration to another. For the CLA, for example, students are typically tested in the first semester of their freshman year and in the last semester of their senior year. The idea is that, by comparing scores across these administrations, various stakeholders can have an idea of how much learning is going on in those years of education. This is an essential aspect of tests of collegiate learning because, as mentioned previously, the entirety of the colleges admissions process amounts to a machine for creating unequal levels of starting ability in incoming college classes. Elite colleges have such onerous admission standards precisely because those colleges are attempting to filter out all but the best-prepared students. Therefore, any attempt to systematically analyze college learning fairly—particularly when attached to high-stakes programs such as Barack Obama's "Race to the Top" proposal to tie federal college aid to assessment has to account for these differences in ability. It's this problem that the CLA's value-added

approach is meant to address. The specific adjustment made by CAE to demonstrate added value is described by Klein et al. 2014:

> Operationally, this adjustment is made by calculating the difference between a school's actual mean CLA score and the mean its students would be expected to earn. For this purpose, the expected mean is derived from the strong empirical relationship between mean CLA and SAT scores across all the schools in the CLA…. For each college, then, we compute the difference between its actual and expected mean CLA score. This difference is called the school's residual score. We make this calculation separately for freshmen and seniors. Next, we compute the difference between the freshmen and seniors' residual scores at the college. Finally, we examine whether the difference between these two residual scores is larger, smaller, or about the same as the difference in residual scores that is typically found at the other colleges in our sample. (424-425)

Klein et. al. admit that there are potential problems with this approach. For example, this type of analysis assumes that they have avoided selection bias—that is, that they are comparing like with like. Systematic differences between the freshman and senior test takers, or between tested students at different schools, would undermine the value-added measurement. Value-added modeling is hampered in this way by the fact that the placement of students into universities is never truly random and that there are always underlying non-random influences that could influence scores. In a piece providing a broad overview of value-added models, Henry I. Braun of ETS notes these difficulties, writing that "it is impossible… to document and model all such irregular

circumstances; yet they may well influence, directly or indirectly, the answers we seek" (10). This problem can be particularly acute with a low-stakes test like the CLA, as variability is introduced not merely in student populations but according to the different sets of students who show up for the test. For example, Banta and Pyke report that at Jamestown College, "During the first year of testing there, the seniors in nursing and a few other majors were not able to take the CLA due to other commitments. The results were disappointing to the faculty. The following year steps were taken to ensure that a more representative sample of seniors was tested, and the results were much improved" (26-27). While anecdotal, this type of story is concerning, in that it reveals how construct-irrelevant variance can impact outcomes. Additionally, the developers admit that it is impossible to say how much of the growth in student scores stems from direct college learning and how much from other factors. "While higher education is a major force in college students' lives," write Klein et. al., "other factors (such as maturation) may have contributed to the improvement in scores between entry and graduation" (426).

Whether these difficulties undermine the usefulness of value-added models entirely is a matter of debate. In 2011 John Ewing, then president of the Mathematics Society of America, published a cutting critique of the popular understanding of value-added models in education. Ewing summarizes the current state of understanding of value-added models in writing, "Value-added modeling pops up everywhere today….Yet most of those promoting value-added modeling are ill-equipped to judge either its effectiveness or its limitations" (Ewing 667). Ewing argues that the use of value-added modeling is a prime example of "mathematical intimidation," which he defines as the attempt to quiet criticism or enforce a particular point of view by treating quantitative

knowledge as more certain, powerful, or true. "As a consequence," Ewing writes, "mathematics that ought to be used to illuminate ends up being used to intimidate" (667). A professional mathematician himself, Ewing is hardly likely to be intimidated. He points out that there have been many challenges to the validity and reliability of value-added modeling. For example, as early as 2003, Daniel McCaffrey, a statistician with the RAND Corporation and a fellow in the American Statistical Association, wrote of value-added modeling, "The research base is currently insufficient to support the use of VAM for high-stakes decisions" (McCaffrey *xx*). Similarly, Ewing quotes an Economics Policy Institute paper which argues that "VAM estimates have proven to be unstable across statistical models, years, and classes" (Baker et. al. 1).

Clearly, there are statistical and empirical issues associated with value-added models. But given the vast differences in student populations across different colleges and universities, a reality that no one involved disputes, some sort of normed comparison across differing populations is necessary. Part of the difficulty for test developers like CAE, and the stakeholders who must interpret standardized test scores, lies in trying to understand a particular school's results relative to other schools, whether in national comparison or in comparison to similar institutions, when the number of CLA schools is relatively small. Hundreds of institutions now participate in the CLA program, but there are some 4,500+ plus degree-granting postsecondary schools in the United States, of which almost 3,000 are 4-year colleges ("Fast Facts"). As with the effort to make the CLA score a meaningful metric for individual students in the eyes of graduate schools and employers, the ability to draw truly meaningful comparisons between institutions likely requires a certain critical mass of participation. Even if such a national context is

created, there are meaningful concerns about how reliable and fair CLA results will be.

Future Directions

The transition from the CLA to the CLA+ is still quite new. In this early stage, we still lack a substantial research literature about the new assessment. Given that CAE carefully controls what information researchers might glean—even administrators at the schools that administer the test, for example, are prohibited from looking at actual student responses—it is unclear how robust the research literature will ever be. Just as importantly, it is far too soon to adjudicate what impact the test will have on individual institutions and on the broader world of American higher education. These impacts will be affected by many factors, certainly including whether the Obama administration's rankings proposal is implemented, how the labor market continues to change in the next decade, and whether institutions are capable of reducing the speed with which tuitions have grown. The CLA, like all tests and assessments, exists in a complex, multivariate context, and that context affects the test and the way it is interpreted.

It's reasonable to expect that the test's mechanisms and sections will continue to evolve, although CAE seems committed to moving forward with the CLA+ as the definitive version for the foreseeable future. The immediate task for CAE is to spread the test to more and more institutions. This challenge will not be easy. Colleges and universities are large human systems, and like most, they evolve slowly, sometimes glacially so. What's more, given the various criticisms of the test, and the significant resources and effort required to implement this kind of assessment, individual institutions may well decide not to adopt it. Some may go with competitors provided by companies like ETS or ACT. It also remains unclear whether most private colleges, particularly elite

schools with prestige, resources, and clout, will feel sufficient pressure to adopt the test. Ultimately, what's needed for researchers and administrators alike is a better grasp of why and how the CLA has been implemented at individual institutions, what challenges they faced in that implementation, and what lessons they drew along the way. Those questions are the subject of the next chapter of this dissertation.

CHAPTER 4: THE TWO CULTURES

In the introductory remarks to the final section of *Writing Assessment in the 21st Century*, titled "Toward a Valid Future," Norbert Elliot and Les Perelman write, "Tension between the composition community and the educational measurement community is a theme that runs through this volume" (407). As the book is a broad overview of the history of writing assessment theory and practice, as well as a consideration of the future of writing assessment, this statement is particularly indicative of a broad and persistent conflict in cultures. From the other side of the divide, Pamela Moss, a professor of education and someone firmly in the educational testing camp, wrote in 1998 that "the field of college writing assessment… appears seriously isolated from the larger educational testing community" (113). Members of both cultures appear, therefore, to agree that there is a divide. In the following chapter, I will discuss the origins and nature of this divide, its stakes, and potentially methods for fixing it. More, I will argue that the divide is less a matter of writing studies scholars in opposition to the field of educational testing, and more a matter of writing studies being a field at war with itself.

A Brief History of Practitioner Writing Assessment

In order to understand the traditional divide between writing practitioners and the educational testing community, it's necessary to undertake a brief history of practitioner

writing assessment. Though the subject itself could fill several books, a broad overview of the development and evolution of how writing teachers, researchers, and administrators assess student writing can be useful for understanding current assessment controversies.

Like the history of college composition as a distinct entity itself, the beginning of college writing assessment is typically dated to the late nineteenth century. As John Brereton writes in his essential history of *The Origins of Composition Studies in the American University* (1995), "The composition course as we know it today, like the university that teaches it, is a product of late nineteenth century America" (40). Our disciplinary histories rarely focus attention on the role of writing assessment in this period. However, assessment was in fact a key aspect of the early development of college composition. As James Berlin notes in *Writing Assessment in Nineteenth-Century American Colleges* (1984), college writing pedagogy was deeply influenced by the entrance exams that were being implemented by many universities of that era. These entrance exams arose out of a perceived lack of prerequisite ability entering American colleges, with a lack of writing skills seen as an area of particular need. These entrance exams were typical of writing assessment for the first half of the 20th century: defined in terms of remediation, seen as lying outside of core writing pedagogy, and frequently instituted on an ad hoc basis.

There was little formal development in writing assessment in the half century that followed those beginnings, in large measure because composition instruction was seen as a service ancillary to the real intellectual work of teaching classics, literature, and theology. Kathleen Blake Yancey's "Looking Back as We Look Forward: Historicizing

Writing Assessment" (1999), one of the most comprehensive and important histories of post-World War II college writing assessment, begins in 1950. While assessment practices did take place prior to this period, Yancey's analysis suggests a rapid increase around that time. The rise of machine calculation and improvements in data analysis that occurred in the post-World War II period likely contributed to this increase in writing assessment practices. In what she notes is a consensus view, Yancey describes three general periods in the history of college writing assessment. In the first period, from 1950-1970, writing assessment came in the form of "objective" tests that were typically multiple-choice affairs that tested grammatical knowledge. In the second, which she dates from 1970-1986, writing assessment was primarily a matter of short timed essays, such as those that persist in the SAT, TOEFL, and GRE, as well as in many placement mechanisms for incoming college students. In the third, from 1986-the present (or the present when she was writing, 1999), writing assessment moved towards programmatic assessment—assessment at the program level rather than at the individual student level—and towards the portfolio assessments that were seen as more authentic and complete than short-answer essays.

During the first period, Yancey writes, "'objective' tests, particularly multiple-choice tests of usage, vocabulary and grammar, dominated practice…. most testing concerns focused on sites ancillary to the classroom" (485). Frequently these tests involved identifying necessary edits to be made in a passage of prewritten text, usually chosen from a set number of potential choices. Sometimes they involved choosing the correct form or syntax from a set of options. These tests were subject to an obvious critique: they did not assess any actual student writing, and so they lacked construct

validity in a social scientific sense. Though they could be demonstrated to have criterion validity, in that they were frequently correlated with deeper evaluations of writing ability, their lack of authentic assessment of student writing ability made them distrusted by students and instructors alike. Part of the reason for their use, despite the gap between the tests and the actual practice of writing, was practical: the GI Bill was democratizing the American university, opening the doors of a college education to students who were outside of the traditional social and economic elite that dominated college populations. "Consequently," Yancey writes, "there were genuine questions as to what to do with these students: where to put them, how and what to teach them, and how to be sure that they learned what they needed" (485). These concerns were legitimate, but they helped cement assessment's status as an external solution to a problem rather than as an integral part of writing pedagogy.

Although they might not have used the term validity, with its social scientific connotations, writing scholars in the next period nevertheless attacked the objective tests as lacking validity. As Huot writes, "The English teaching profession vociferously protested English and writing tests that contained no writing" (24). This natural aversion to inauthentic or invalid assessment led to Yancey's second period and the development of timed essay tests, such as those that continue to be used in standardized tests and which are often still involved in the placement of college students into different levels of introductory writing courses. Yancey names Edward White, then as now a major figure in the theory and practice of writing assessment, as a major figure in the move away from objective tests and towards timed essays. In his role as an administrator in the California State University system, White spearheaded timed essay assessments that, given the size

of that system, were taken by thousands of students. Yancey describes the basic requirements of this kind of testing, which was inspired in large measure by ETS's AP exams:

> a classroom-implemented curriculum culminating in a final essay test that met adequate psychometric reliability standards through several quite explicit procedures: (1) using writing "prompts" that directed students; (2) selecting "anchor" papers and scoring guides that directed teacher-readers who rated; and (3) devising methods of calculating "acceptable" agreement. (490)

A common complaint about such tests has been that the standardized prompts restrict the freedom to direct one's own writing in a way that is common to many writing classes. Years after he helped pioneer such assessments in the CSU system, White defended the use of standardized prompts by pointing out that in both their college lives and their post-college lives, most students will write under similar direction and constraints. "The demand to write, in school no less than on the job," writes White, "is almost always an external demand, and an exacting one" (57). In other words, the fact that students do not get to choose what they write about on essay tests makes them more like most real-world writing situations, not less.

Complaints about the inauthenticity or unfairness of timed essay tests persisted. Then, as now, writing scholars feared that timed essay assignments, almost exclusively utilizing prompts calling for students to write unresearched opinion essays, which did not fairly or fully represent an individual student's writing abilities. College writing, after all, requires a diverse array of skills that are employed to satisfy several different assignment types. Students whose composition skills lie outside of the short, persuasive essays that

predominate in standardized writing assessments might be unduly disadvantaged by this type of testing instrument. White, though a qualified defender of standardized essay tests, admits that if a student is inexperienced or misinterprets the prompt, the resulting essay is likely to be "muddled, error-ridden, inappropriate, or just bad" (53). Expressed more positively, Yancey defines the attitude underlying calls for portfolio assessment: "if one text increases the validity of a test, how much more so two or three texts?" (491). This is one of the simplest virtues of a portfolio system: the expansion of the amount of data to be assessed. This virtue is matched, of course, with the added burden of more work for portfolio reviewers.

In the mid-1980s, a new portfolio system was pioneered by Peter Elbow and Pat Belanoff, then of SUNY Stony Brook. The system they developed included many of the hallmarks of portfolio assessment that endure to this day: texts drawn from classroom work that were then revised by the students; a variety of tasks, prompts, and assignments reflected in the portfolio; and a dichotomous, pass/fail system of scoring that was arrived at through the mutual agreement of multiple raters, a consensus-based approach that allowed for talking out disputes between raters. Since this influential work, there have been many attempts to implement portfolio assessments in various contexts, changing the Elbow and Belanoff system as needed. Writing ten years later, Donald Daiker, Jeff Sommers, and Gail Stygall list five common features most portfolio assessments share:

Most portfolios

- include multiple samples of writing from a number of occasions;
- require a variety of kinds of genres of writing;

- provide opportunities for revision and request evidence of the revision process;
- ask students for their reflections—on their portfolio, on their writing process or history, or on themselves as writers; and
- offer important choices to the writer. (257)

Portfolio assessments are now in use at a variety of educational institutions and for a variety of purposes. However, there are practical reasons that they remain a comparatively small part of the broader world of writing assessment. This practical difficulty is obvious: the resources required to implement such portfolio systems are considerable, both in terms of incorporating them into classes and in terms of the actual assessment. Because these systems involve revising classroom texts and incorporating them into the portfolio, time in class must be devoted to their assembly. As Daiker, Sommers, and Stygall write, for "busy faculty members… even if they believe that portfolios may be useful for teaching and assessment, the time required to develop assignments and read the students' writings would be too great" (277). This is a particularly acute problem for larger programs. A writing program that offers a dozen freshman composition classes a semester might be able to effect the consensus, dialogic assessment of portfolios that Elbow and Belanoff advocated. But at a school like Purdue, where sections of Introductory Composition in a given semester always number in the hundreds, this task becomes monumental, and likely entirely unworkable.

Further, in their emphasis on local definitions of success, their tendency to eschew strict rubrics, and their tendency to include different types of texts and assignments from student to student, portfolio assessments cut directly against many of the basic

assumptions of conventional educational testing. Reliability concerns are a persistent aspect of portfolio discussion. For example, a 1993 study discussing internal research at the University of Wisconsin found that "Reliability estimates were low to moderate" for portfolio scoring (Nystrand, Cohen, and Dowling 53), although the article discussed some possible reforms that might improve reliability. That means that students may be rated as proficient or not depending on which individual rater(s) evaluate their portfolio. This potential inconsistency is clearly suboptimal from the standpoint of basic fairness. What's more, it introduces uncertainty and imprecision into the broader system of assessment and credentialing that are an essential part of the contemporary university and its place within the economy. The desire among writing instructors for a more authentic, more comprehensive, deeper system of assessment pits them against the perceived labor-market function of higher education.

In this sense, portfolios contribute directly to the central tension within this chapter: the frequently conflicting cultures of writing researchers and the developers of standardized tests.

Sources of Friction

As previously stated, the evolution of collegiate writing assessment from multiple-choice tests to timed essays to portfolio systems can be seen as a gradual movement from privileging reliability to privileging validity. However, in most testing circles, these goals are both seen as essential elements of effective and responsible assessment. While some in the educational testing world would concede that test development necessarily entails tradeoffs in validity and reliability, almost all would argue that both must be present for an assessment to deliver useful information. Writing

instructors and administrators, in contrast, have tended to be less concerned with these traditional aspects of social scientific research. However, as strides have been made to establish assessment as a rich and valuable aspect of writing theory, more and more writing researchers have attempted to use the vocabulary of testing to articulate the superiority of certain types of assessment. This adoption remains partial and contested.

Portfolio assessments are highly indicative of the fitful adoption of the language and perspectives of educational testing community. From an intuitive point of view, portfolios seem to improve the validity of writing assessment relative to objective tests or timed essays. Simply collecting more evidence would seem to positively impact validity, and that is particularly true if what we intend to measure is the ability to succeed at the broad types of writing employed in college. As Daiker, Sommers, and Stygall write, "writing is a complex, multifaceted activity that cannot be appropriately represented by a single genre: not by exposition, not by argument, not by critical analysis" (257). That their breadth makes portfolios more valid is a view widely held in the writing assessment community, although Huot cites scholars like Samuel Messick and Lorrie Shephard in arguing that this intuitive sense of validity is undertheorized (49-50). However, this increase in validity comes at the previously-noted cost of reliability. Yancey admits, despite being a supporter of portfolios, that "portfolios are 'messy'—that is, they are composed of multiple kinds of texts, and different students compose quite different portfolios, even in the same setting and for the same purposes, which in turn can make evaluating them difficult" (493). That difference between what is being assessed from student to student is precisely what reliability procedures are meant to avoid, and this

diversity in student responses seems to cut directly against the instincts of the psychometric and educational testing communities.

Though the long-term trend within writing assessments that are controlled by writing programs and professors has been from more reliable but less valid instruments to more valid but less reliable, the trend from “inauthentic” to “authentic” tests is not universal or uncomplicated. White, likely the most influential scholar in the history of writing assessment, has argued that writing assessment needs to focus on reliability for issues of simple fairness, writing that "Reliability is a simple way of talking about fairness to test takers, and if we are not interested in fairness, we have no business giving tests or using test results” (“Holistic” 93). Generally, though, the push towards validity at the expense of reliability is indicative of the “two cultures” referenced earlier in this text. What’s more, while writing researchers often invoke the concept of “construct validity” – the notion that validity primarily entails deciding whether an assessment actually measures what it was intended to measure—it will likely become important for writing researchers to engage with more complex notions of validity. For example, concurrent validity, which involves comparing results on one type of assessment to results on another, to see whether they may cross-validate each other, is common in educational testing circles.

One of the central sources of conflict between these groups is the tension between state and national standards and the desire for local control. As discussed in Chapter Two, assessment is a major part of a national effort to reform university education. This movement has been the target of considerable criticism from scholars within the university, much of it fair and legitimate. A particular fear for instructors is the loss of

local control over their pedagogy and their grading, legitimate fears in a country that has recently experienced the introduction of No Child Left Behind and the Common Core. While the ultimate strengths and weaknesses of these policies are ultimately subject to considerable debate, there is little question that both, to a degree, restrict teacher autonomy and control of curriculum. Here at Purdue, the fear that professors would lose control over their pedagogy and grading was sufficient that President Daniels felt compelled to reassure the faculty that the CLA+ would never be used to replace traditional grading (see Chapter 5). The commitment to the local is a cherished theoretical and political commitment of many writing scholars. Writing in a February 2014 *College Composition and Communication* review essay that concerned assessment, Northeastern University professor and assessment expert Chris Gallagher sums up this attitude in writing

> all writing assessment is local. This proposition does not suggest that compositionists are unaware of state, national, and international assessments or indifferent to forces operating at these levels. Rather, it posits that assessment decisions are always experienced locally—by people in the places they teach and learn. It also insists that the construct being assessed—writing—is itself a highly contextualized activity, learned and practiced by individuals and groups in specific rhetorical situations—and so assessment of it must be, too. Not least, the proposition is axiological as well as ontological: a declaration that writing assessment must be conducted by those who know something about writing and who live most directly with the consequences of assessments. (487-488)

This is a reasonable philosophy, but one that must be balanced with a frank admission of the inevitability of state and national standards. Writing programs, like all educational endeavors, are embedded in institutions, and those institutions are parts of systems of power. No educators exert total control over their local contexts. The key must be to defend local control effectively, in part by understanding and working with systems and requirements that are enforced from above. As Maurice Scharton writes, "one must accept reasonable limits on home rule in the classroom. Society… has a legitimate interest in what one does in one's classroom" (61).

Beyond the specific empirical and theoretical divisions, a persistent divide in what we might inexactly refer to as culture contributes to the lack of cooperation between these groups. Few who are informed about issues within writing assessment doubt that such cultural tension exists. As Scharton writes, "however we rationalize assessment products, we cannot avoid the sad realization that assessments define opposing political camps" (54). As Huot puts it, literature on writing assessment produced within the rhetoric and composition community frequently casts writing scholars as "combatants who wrestle away control of writing assessment from testing companies who would ignore the need for writing assessment even to include any student writing" (36). As Huot argues, the reality is far more complex, and no inherent reason would keep compositionists and test developers from working with mutual respect, even if we concede that their differing assumptions and values will frequently provoke disagreement. But a considerable distrust between these communities clearly exists. As Keith Rhodes and Monica McFawn Robinson write in a 2013 article, assessment efforts that cannot be aggregated with other data or removed from their individual contexts are not "appealing to anyone beyond the

relatively small circle of those already immersed in composition scholarship" (16). If writing professors and administrators are to respond constructively to the challenge and opportunity that assessments like the CLA+ represent, work must be done to bridge this gap. This work is not merely valuable as a means to bring researchers from different perspectives closer together. It is an essential part of ensuring that scholars in writing studies retain administrative power. As Scharton points out, when writing instructors do not engage on issues of importance in standardized assessment, "the English profession suffers a corresponding loss of credibility among the powerful people who do not share its orthodoxies" (60).

The Higher Education Assessment Movement and the Two Cultures

Scharton argues, then, that the field of English endangers itself by refusing to engage with the techniques, philosophies, and research of the educational testing community. The current political movement to develop assessments of higher education reveals this danger aptly. In this case, the "powerful people who do not share its orthodoxies" potentially includes the politicians, such as those in the Bush and Obama administrations, who are currently advocating for more "objective" assessments of student learning; many members of commissions, panels, and committees who develop analyses and recommendations for those politicians; and the federal, state, and local-level administrators, including college administrators, who actually implement policy. This assessment might seem bleak, yet little question remains that the education testing community, and in particular the major testing companies and nonprofits, are driving the current state of assessment to a greater degree than writing researchers and instructors.

Multiple factors contribute to this predominance, and many of them lay outside of the hands of writing and English faculty. Corporate, political, and non-profit interests have tremendous power over educational policy and practices in this country. Consider, for example, the Common Core curriculum. Originally developed in 2009 by a panel commissioned by the National Governor's Association, the standards were be "research and evidence-based, internationally benchmarked, aligned with college and work expectations and include rigorous content and skills" ("Forty-Nine States and Territories"). By creating a set of national standards in mathematics and the language arts, and creating incentives for applying those standards like the Obama administration's Race to the Top initiative, the Common Core could become one of the most significant policy evolutions in the history of American education. The speed with which the Core standards were adopted by various state legislatures was remarkable, given the profound nature of the change, and would come to attract considerable controversy. At issue in particular was the influence of the Bill and Melinda Gates Foundation, the powerhouse nonprofit organization that is funded by billions of dollars of charitable contributions from the Gates family. As this controversy bloomed in 2014, an investigative piece from *the Washington Post* by Lindsey Layton demonstrated the degree to which Gates had personally impacted the Common Core push. "The Bill and Melinda Gates Foundation didn't just bankroll the development of what became known as the Common Core State Standards," wrote Layton. "With more than $200 million, the foundation also built political support across the country, persuading state governments to make systemic and costly changes" ("How Bill Gates Pulled Off"). Although the article did not allege

explicit corruption, it detailed how far Gates Foundation money went in persuading various politicians, experts, and stakeholders to support the Core.

By setting the educational agenda for K-12, meanwhile, the Core necessarily impacts college pedagogy. Discussions of college composition pedagogy often include claims that students from high school arrive unprepared for their college writing tasks. For example, the National Center for Public Policy and Higher Education argues in a position paper that "while many states have made progress in getting more students to take the high school courses necessary for college readiness… only a few have specified an explicit set of performance skills in reading, writing, and math that signify college readiness" ("Beyond the Rhetoric"). The Common Core's exact requirements will go a long way towards determining those required performance skills, which will in turn play a role in whether students arrive on our campuses ready to succeed in their writing classes or not.

A consideration of the makeup of the major presidential educational commissions is also illustrative (see Chapter Two). For example, President Reagan's National Commission on Excellence in Education, the authors of *A Nation at Risk*, included five members of local, state, and national school boards; four university presidents; three principals and superintendents; two members of industry; two members of disciplinary boards; a professor of chemistry and a professor of physics; and a high school teacher. George W. Bush's Commission on the Future of Higher Education included five members of industry; four current or former presidents of colleges and universities; four members of various educational boards or trusts; and five professors. Not one member of these two commissions that have done so much to dictate recent higher education policy

had as his or her primary research interest writing, composition, rhetoric, or any similar subject that could be plausibly considered a part of the world of writing studies. The issue of cause and effect is cloudy here; it's unclear if, for example, these presidential commissions excluded writing scholars precisely because of the general resistance of writing scholars to assessment and quantitative methods. But one way or another, writing studies as a field of research and pedagogy had no voice in these important commissions.

Clearly, then, there is a degree to which the persistence of the two cultures dynamic lies outside of the control of writing practitioners. Political and corporate interests have worked to remake education without the input of writing teachers and researchers, as exemplified by the power of the Gates Foundation to enact the Common Core standards. Writing instructors themselves frequently find themselves marginalized in the development of writing standards, and this marginalization naturally leads to feelings of skepticism and distrust that perpetuate the cycle. But this marginalization should not allow us to excuse the ways in which scholars from English and writing have essentially self-selected themselves outside of the conversation. As Huot, Scharton, Moss, and others have noted, too often scholars from within the broad world of writing studies have self-marginalized, fearful of being seen as taking part in the legitimation of hegemonic power structures. The broader question is why. Why have so many within the field of writing studies have resisted taking part in these conversations? Why are so many on this side of the cultural divide unable or unwilling to take part in debates that have obvious and considerable consequences for the field? The answer has much to do with the contested role of empirical research generally and quantitative research specifically in the world of writing research.

The Contested Role of Quantification in Writing Studies

Writing studies, and its affiliated field rhetoric and composition, has had a long and tangled relationship with empirical research generally and quantitative research in particular. The complexity of this relationship has a significant impact on the role of writing instructors in the development of standardized assessments of writing. Because many scholars within the broad world of writing research have argued that there is little or no place therein for quantification or the techniques of the social sciences, few graduate students and young professors learn these techniques. That in turn limits the ways in which members of the field can impact debates about standardized testing. Because quantification and social science theory are so deeply entrenched in standardized testing and large-scale educational assessment, a field that refuses to use them will necessarily find itself on the outside looking in when it comes time to assess.

That rhetoric and composition is generally not welcoming of quantitative research has been a commonplace understanding for several decades. A series of influential articles identified the dearth of empirical research within the field in recent decades. As early as 1996, Davida Charney reported in her article "Empiricism is Not a Four-Letter Word" that a debate was raging about "whether empirical methods have any legitimate place in composition studies" (567). The directness of that statement helps to demonstrate the intensity of the resistance to these ways of knowing. It is certainly true that writing studies, as a subfield within both English specifically and the liberal arts more generally, could be expected to embrace more humanistic types of research methods such as close reading and theory. But to question the appropriateness of empirical methods writ large is a stark statement. Nearly a decade later in 2005, Richard Haswell echoed Charney's

statements, in his article "NCTE/CCCC's Recent War on Scholarship." Haswell's title suggests the polemical nature of his argument, indicting both the National Council of Teachers of English, the primary professional organization of rhetoric and composition, and the Conference on College Composition and Communication, the field's most prominent conference, for resistance to empirical research. Haswell demonstrates that what he calls "RAD" research – replicable, aggregable, data-driven—has become remarkably underrepresented among the three major journals published by the NCTE, *College Composition and Communication*, *College English*, and *Research in the Teaching of English*. Given that these journals are considered extremely prestigious, this dearth of publication sends a clear signal to writing scholars that this type of research is not valued. Describing internal resistance to this kind of scholarship as a "silent, internecine, self-destructive war" (199), Haswell argues that by abdicating these types of research methods, writing researchers lose the ability to influence essential parts of education policy, the "ability to deflect outside criticism with solid and ever-strengthening data" (219). In a 2013, Rhodes and McFawn Robinson could still report that "while some scholarship has clearly turned away from social construction in recent years, we believe that its influence continues—most obviously in the durable arguments against the 'positivism' of data collection" (8). Rhodes and McFawn Robinson go on to describe a field of writing researchers who still struggle to develop a meaningful set of shared knowledge, thanks to the prohibition against systemizing methods that Charney and Haswell discuss.

Understanding this dynamic, and the dangers presented by it, requires recognizing that it is a fairly recent development in the history of the field. It would be easy to assume

that quantitative and empirical research have always been strangers to rhetoric and composition, given that the field emerged primarily from English departments and that its scholars have often been those initially trained in literature scholarship. But in fact, empirical research has a long and noble tradition within the field of writing research. Although the notorious difficulty in deciding when writing studies truly began as an academic discipline complicates the discussion, there is little doubt that empirical research into student writing practices have existed for at least as long as the field itself. In his landmark book *The Making of Knowledge in Composition* (1987), Steven North looks back as far as the early 1960s, arguing that the "literacy crisis" that was percolating in the American media at the time helped create the modern field of composition. In that history, North identifies four major schools within writing research: experimentalists, clinicians, formalists, and ethnographers. However much we might agree or disagree with this taxonomy, North's ability to sub-divide empirical researchers within writing studies in this way demonstrates the existence of a robust, varied set of subjects and methodologies. At the time of North's writing, that diversity in methods and acceptance of empirical research was secure enough for Janice Lauer and J. William Asher to publish *Composition Research: Empirical Designs*, a handbook for performing empirical research, the following year. Yet even as that text was being published, the tide was turning against empirical writing research.

The change in fortunes for empirical writing research has generally been ascribed to the rise of cultural studies as the dominant theoretical framework of writing research. Scholars such as James Berlin, Elizabeth Flynn, Carl Herndl, and many others argued that the purpose of writing scholarship should be emancipatory, that conventional writing

classrooms were sites where students could be taught to resist hegemonic power relations, and that empirical research (and especially quantitative research) was necessarily the domain of establishment power. Though given a variety of names, including not only cultural studies but critical studies, emancipatory pedagogy, critical pedagogy, and similar, the broad trend has been unmistakable in the last several decades of composition research. Berlin, perhaps the most essential to this change, co-edited a 1992 volume titled *Cultural Studies in the English Classroom* that has frequently been identified as a breakthrough for these theories in composition. Berlin argued that the role of writing pedagogy should not be merely to teach students to describe culture in writing but to understand how they are complicit in traditional power structures within culture and, in their writing, oppose them. In 1993, Herndl's article "Teaching Discourse and Reproducing Culture" argued the by-now common view that college writing research frequently acted to reify and solidify existing power relations, and that attempts to systematize or formalize our inquiry were especially guilty in this regard. In 1995, Flynn's "Feminism and Scientism" expressed a point of view that would become similarly commonplace within the field, that "[f]eminist critiques of the sciences and the social sciences have also made evident the dangers inherent in identifications with fields that have traditionally been male-dominated and valorize epistemologies that endanger those in marginalized positions" (355).

Whatever the exact origins of composition's embrace of cultural studies, by the turn of the 21st-century, the denigration of quantitative research as politically, theoretically, or practically unhelpful was an assumed part of the landscape of composition. In a 2001 essay in an edited collection, John Trimbur and Diana George

would write that "cultural studies has insinuated itself into the mainstream of composition" (71). By 2005, Richard Fulkerson could write regarding the dominant "social turn" in composition studies that "in point of fact, virtually no one in contemporary composition theory assumes any epistemology other than a vaguely interactionist constructivism. We have rejected quantification and any attempts to reach Truth about our business by scientific means" (662). Fulkerson argues that these epistemological assumptions "determine what sort of scholarly research is acceptable as grounding" for approaches to writing pedagogy and "also control what students are taught regarding 'proof' in their own reading and writing" (662). In other words, these assumptions dictate both pedagogical practice and research methods, leaving us as field with only "sophisticated lore."

The problems with resistance to empiricism and quantification by people within writing studies are multiple. First, as this chapter has argued, it amounts to self-marginalization within certain discourse communities. While many scholars within writing studies have argued for non-quantitative theories and practices of assessment, and have had some limited success in site-specific reforms, for the most part the refusal to take part in quantitative research amounts to a refusal to take part in serious debates about assessment. While most scholars in writing studies probably lament the current preeminence of quantitative ways of knowing, there is little reason to believe that this preeminence will fade in the near future, and a refusal to advocate for our values in those terms will result in an inability to help shape the future of pedagogy and policy. "Without a full philosophical shift," write Rhodes and McFawn Robinson, writing studies will not "be likely to persuade more realist or idealist audiences that it has anything to offer to

anyone outside its circle" (10). This absence of influence can be seen as a failure to properly take stock of the rhetorical situation: a key element of rhetoric has always been recognizing the context in which you argue. For a field full of rhetoricians to fail to recognize the self-marginalization inherent in their refusal to quantify demonstrates the profound head-in-the-sand quality of the current condition.

Second, the refusal to quantify leaves writing scholars, and writing programs, unable to defend themselves when their principles and practices are challenged. Since our programs are embedded in institutional, political, and economic contexts, they must be ready to respond to challenges that take many forms. The exigency of context is the source of White's famous White's Law: assess or be assessed. The kinds of inquiry into our teaching and administrative practices that we refuse to do out of theoretical resistance leave us vulnerable to critique in that area. And with so much of the world of policy and administration now taken with the notion that arguments involving numbers and statistics, our refusal to quantify represents a glaring vulnerability indeed. Haswell advocates for what he calls "anticipatory numbering," or undertaking quantitative self-inquiry as a means of preempting quantitative review from outside forces. If "compositionists can analyze numbering as a rhetorical commerce," writes Haswell, "they can adapt this commerce as an argument for one of their own causes, the championing of local over outside assessment of writing" ("Fighting Number with Number" 414). Note that Haswell frames this work not as an inevitable capitulation to the primacy of the number, but as a way to defend our values and preferences through the skillful deployment of numbers. This deployment need not be a universal or even common aspect of the pedagogical, administrative, and research practices of our field. Rather, a relatively small number of

writing researchers and administrators could work effectively to undertake quantitative research and program assessment in order to undertake the anticipatory numbering Haswell advocates.

Of course, in order to undertake such work, scholars must first be adequately trained in it, and this represents a not-inconsiderable amount of effort for the field. Because so few writing researchers have been working with numbers and systems of formal empirical research methodologies in the recent past, it's likely that a correspondingly small number of professors at graduate programs in the field feel qualified to teach students to use them. This lack of experienced professors potentially becomes a vicious cycle in which the inability to effectively utilize a set of techniques as essential to the modern research university as quantitative methods is passed down from one generation of writing scholars to the next. This problem could be ameliorated in a number of ways, however. First, there are an abundant number of books and manuals devoted to research methods and statistics, including many that are explicitly for beginners. Second, graduate students housed in large research universities (the type most likely to send graduates into programs where they will train graduate students in turn) typically have the ability to take courses from other fields such as education, psychology, and statistics, where they could take courses in assessment and research methods, allowing them to bring these techniques home to their own departments. Some traditionalist rhetoric and composition scholars might bemoan the time and effort invested in these courses; student attention, after all, is a finite resource, particularly when viewed from the standpoint of a brief two or three years of PhD coursework. But it would not require a large number of scholars in writing studies pursuing quantitative

literacy to effectively disseminate broader knowledge of quantitative methods into the field. There is room for theory, rhetoric, pedagogy, qualitative, and quantitative inquiry alike.

In the years since critiques like those of Charney and Haswell, there have been numerous claims that rhetoric and composition is ready to broaden its methods again and embrace empirical research. For example, in September of 2012, *College Composition and Communication*, widely considered the field's flagship journal, ran a special issue dedicated to broadening rhetoric and composition's research methods. In this issue, alongside articles on more conventional approaches in rhetoric and composition such as archival research and discourse analysis, ran articles on eye tracking, data mining, and graphing of large data sets. "It's a truism that we have more information than the world has ever seen," writes editor Kathleen Blake Yancey in the issue's introduction, and the work ahead requires us in rhetoric and composition "to begin to make meaning with it, especially in contexts calling for research that is replicable, aggregable, and data-supported" ("From the Editor" 11). This call echoes those of Haswell and Charney very well, and it reflects agreement that the field badly needs to develop a shared body of knowledge, one that utilizes consistent, systemized methods that allow researchers in different contexts to be intelligible to each other. Speaking anecdotally, as a graduate student with a keen interest in empiricism and quantitative methods, I have been counseled by mentors in the field that more empirical work is a necessity for our disciplinary health.

Still, the turn back towards empiricism and quantification remains more theoretical than actual. For example, in the ten issues of *College Composition and*

Communication published since that special issue, not a single article that could be considered a work of quantitative empirical research has appeared. The condition is similar in the other major NCTE journals, such as *College English,* or even the most empirically-minded of these journals, *Research in the Teaching of English*, which ten years after Fulkerson's article still "publishes primarily ethnographic studies" as he wrote in 2005 (662). The Conference on College Composition and Communication occasionally hosts panels on the need to quantify, such as 2014's H.19, "Collecting, Analyzing, and Talking about Data," or my own presentation, "Statistical Hands, Rhetorical Hearts." But reviewing the program of that year's panels reveals no panels that actually identify themselves as presenting quantitative research directly. In other words, better than a decade after Haswell's article and almost two since Charney's, writing studies still contains more debate and discussion *about* quantitative methods than research *utilizing* quantitative methods. In a world where assessment is primarily considered a quantitative enterprise, this lack of quantitative research only leaves writing scholars further out of the picture.

The Road Ahead: Reasons for Optimism?

Still, there is some reason for optimism that the work of healing the rift between the two cultures is progressing. Journals like *Assessing Writing* and *The Journal of Writing Assessment* publish deep considerations of assessment theory and practice, much of it coming from writing studies scholars. Important journals that focus on administrative and pedagogical practice within writing programs, such as the *Writing Center Journal* and *Writing Programs Administration Journal* regularly discuss assessment, demonstrating the degree to which these issues are of continuing importance

to the conduct of actual writing programs, even if the most prestigious of our journals still largely ignore these practices.

Whatever the strengths and weaknesses of current models, there is little question that great strides have been made in the practical application of writing assessment techniques in the last several decades. This growth suggests why writing scholars might be naturally skeptical about an assessment mechanism like the CLA+. Having developed a real theoretical and practical literature for writing assessment in the last 25 years, writing scholars have legitimate reasons to worry about the imposition of standardized assessments on their pedagogy. It's for exactly these reasons that writing scholars must be willing to grapple with the terms and ideas of educational assessment: to ensure that this growth has not been in vain. Yes, the field of educational testing must bend as well, and should show proper respect to writing scholars by listening to and evaluating our arguments. But the refusal to engage them, or to ever try and speak in their terms, is a rhetorical failure that has potentially profound negative consequences for the field.

In this sense, the fight I have described in this chapter is a fight within the field of writing studies as much as it is a fight against educational testing. Rhetoric and composition is notoriously a field of constant self-examination, one which frequently asks what defines the discipline and what its place is within the contemporary university. Continued resistance to empiricism and enumeration generally and quantitative writing assessment specifically reflect the field's various research domains and theoretical commitments cutting against each other. On the one hand, the field has a set of theoretical and political commitments, described in this chapter, that articulate a principled resistance to the rule of quantitative knowledge that is common to many parts

of the contemporary world. On the other hand, the field wants disciplinary control of aspects of writing education like assessment, which often requires the use of numbers, particularly if that assessment is to speak to various stakeholders who do not share our assumptions. Arguments like that of Charney, Haswell, and MacDonald reflect the recognition that if we are to truly influence policy efforts like the assessment push here at Purdue, we must recognize the rhetorical context and speak in the language of our audience, according to our purpose. This is not in any sense to reject out of hand the concerns of critics who view these techniques with suspicion. Rather, we should view the role of quantitative assessment as a means through which to protect the traditional values those critics defend.

The self-same text I quoted from at the beginning of this chapter to acknowledge the existence of this broad cultural divide, *Writing Assessment in the 21st Century*, features contributions from thirty compositionists and five employees from ETS. This kind of collaboration should become, if not the norm, then a far more common feature of writing scholarship on assessment. We can't afford not to listen to these voices while we advocate for our own values and techniques. Though the book is frank in documenting traditional disagreements, the book offers, according to its editors, "evidence of a narrowing gap between these two communities" (13), and was produced for the explicit purpose of bringing those from these different perspectives closer together. The text reflects a growing acknowledgment from writing programs administrators that the pressure on such programs to demonstrate the value of their teaching is not going away. Only by recognizing these pressures and grappling with the ideas and techniques of the educational testing community can we ensure that we respond effectively, in a way that

ensures we can protect our teaching and our values. This dissertation is intended as a part of that effort.

CHAPTER FIVE: LOCAL CONTEXT, LOCAL CONTROVERSY

Although the research literature on higher education policy is vast and varied, depictions of on-the-ground realities of administrative work at American colleges and universities is limited. The most likely reason for this limitation is the frequent opacity and secrecy of administrative affairs, even at public universities that ostensibly have a responsibility to make their inner workings subject to public review. Discussions of higher education policy also tend towards national perspectives due to the desire for one's research to appear relevant to a wide audience. Work that is too locally focused might appear to be limited or uninteresting, which could have professional and academic consequences. Finally, those who are inclined to investigate local histories might find that they lack the methodological and practical skills necessary to undertake that kind of investigation, given that finding necessary information likely requires interviewing and requests for internal documentation, rather than simply through accessing publicly-available texts. Whatever the reasons, the bias towards the bird's-eye view in higher education policy research is clear.

This lack of local histories risks leaving us with an incomplete picture of how large-scale initiatives like the higher education assessment movement actually come to fruition on the ground. While major educational movements often begin at the highest echelons of our politics and policy world, all implementation of these movements is local, as it is in institutions that these changes actually take place. What is the relationship

between the rhetoric and ideals of the higher education assessment movement and the actual practice of assessment in individual schools and systems? What changes along the path from national agitation for these reforms to the local practices that they produce? How do institutions choose assessment instruments to measure the effectiveness of their teaching? What types of resistance and controversy occur when individual colleges and universities attempt to introduce standardized assessments like the CLA+ into their regular practice? This chapter attempts to provide partial answers for these questions by detailing the history of the assessment effort at Purdue University, discussing its roots in an administrative change in the university and the still-contested role of the CLA+ in the future of Purdue's undergraduate education.

Local Contexts

Purdue University is a large, Midwestern public university system, with its flagship campus located in West Lafayette, Indiana. With over 40,000 students matriculated in undergraduate and graduate programs, Purdue—West Lafayette is close in size to its sister school and rival, Indiana University in Bloomington. Purdue is a member of the Big Ten athletic conference and the Committee on Institutional Cooperation, an educational consortium of the Big Ten universities and the University of Chicago, which was once a Big Ten member. Purdue is defined as a "RU/VH," or very high research activity university, by the Carnegie Foundation for the Advancement of Teaching ("Purdue University-Main Campus"). Purdue enjoys a strong international reputation, frequently appearing on lists of the most prestigious research universities in the country. In the 2014 *US News and World Report* college rankings, the most influential of such lists, Purdue was ranked #62 out of national universities, tied with

schools such as Brigham Young University and the University of Maryland – College Park ("National University Rankings").

Founded in 1869, Purdue is a land grant college, chartered under the auspices of the Morrill Act, a piece of Congressional legislation from 1862 that authorized the creation and funding of universities on what was then the frontier. The official name of the legislation, "An Act Donating Public Lands to the Several States and Territories which may provide Colleges for the Benefit of Agriculture and the Mechanic Arts," describes both the reason for the term "land grant college" and the conditions under which that land was given. Named for Vermont Congressman Justin Morrill who sponsored it, the Morrill Act allowed for the donation of federal land for the purpose of starting colleges. Such legislation was necessary because of a perceived need for more universities in the western parts of the country, part of a larger push by the federal government to spur settlement of these areas. Specifically, the Act tasked administrators of new colleges "without excluding other scientific and classical studies and including military tactic, to teach such branches of learning as are related to agriculture and the mechanic arts, in such manner as the legislatures of the States may respectively prescribe, in order to promote the liberal and practical education of the industrial classes" ("Morrill Act").

Today, Purdue is a respected research institution, known especially for its Engineering, Computer Science, and Agriculture programs. In particular, the Aeronautic and Astronautic Engineering program is one of the most competitive in the world, with a long tradition of graduating students who have gone on to work for NASA, including Neil Armstrong. Purdue is also known for its large international student population. As of

Fall 2014, Purdue was home to more than 8,700 international students from 125 countries, ranking second in percentage of international students among public universities ("International Students and Scholars"). In addition to its main campus in West Lafayette, the Purdue system includes Purdue Calumet, Purdue North-Central, and two schools in the joint Indiana University—Purdue University system. In recent years, Purdue has dramatically expanded its physical facilities, including its dorms, gyms, and dining halls, and a large research park where much cutting-edge research is performed in fields such as nanotechnology and genetics. Such expansions of physical infrastructure tend to attract competitive undergraduate students, which in turn improves a college's placement in the aforementioned rankings. Whether the expansion of facilities and increased ability to attract highly sought-after students actually contributes to the core educational mission of a university is a separate question.

As noted above, Purdue enjoys a strong academic reputation, and its ability to attract competitive international students ranks with some of the most competitive public universities in the country. Additionally, Purdue's well-known STEM focus places it in a place of particular prestige in an era when STEM education is frequently considered more financially desirable than education in other fields. But as the recent push for greater empirical assessment of American colleges has demonstrated, reputation is not always a valid indicator of quality. Increasingly, external assessments of quality are seen as necessary. The most common such assessments remain the accreditation process.

Previous Assessment: Accreditation

Accreditation agencies play an essential role in the assessment of any college or university. The US Department of Education defines the purpose of accreditation as

"ensur[ing] that education provided by institutions of higher education meets acceptable levels of quality" ("Accreditation in the United States"). Towards that end, the federal government has identified some 15 national and regional agencies that can accredit institutions of higher learning, along with many specialized accrediting agencies that are dedicated to particular academic fields and degrees. These agencies are "private educational associations of regional or national scope [which] develop evaluation criteria and conduct peer evaluations to assess whether or not those criteria are met" ("Overview of Accreditation"). Purdue University is accredited by the Higher Learning Commission of the North Central Association of Colleges and Schools, an agency that accredits hundreds of schools including fellow Big Ten universities such as the University of Wisconsin-Madison, the University of Nebraska-Lincoln, and Indiana University-Bloomington.

Purdue was first accredited in 1913, and underwent its last full, external accreditation review in the 1999–2000 school year ("Purdue University Accreditation"). Guided by a 12-person steering committee, accrediting officials from the North Central Association (NCA) gathered data from the university over a number of months. Five NCA study committees examined key aspects of the university and reported back. The final report, which numbered some 343 pages, made five core claims about the institutional performance and health of Purdue University in the year 2000:

1. The institution has clear and publicly stated purposes consistent with its mission and appropriate to an institution of higher education.
2. The institution has effectively organized the human, financial, and physical resources necessary to accomplish its purposes.

3. The institution is accomplishing its educational and other purposes.
4. The institution can continue to accomplish its purposes and strengthen its educational effectiveness.
5. The institution demonstrates integrity in its practices and relationships. ("Accreditation of Purdue University's West Lafayette Campus")

An additional self-study was undertaken from 2008 to 2009, with a full report published in 2010. The Higher Learning Commission accepted this self-study as sufficient evidence for renewal of accreditation for an additional decade. The 2010 report was compiled by a 14-member evaluation team of scholars and university personnel from other institutions, who were provided information by liaisons from each major-granting department and program within Purdue. The evaluation team unanimously advised continuing accreditation.

While accreditation remains a key aspect of proving collegiate effectiveness, accreditation itself is insufficient for the kind of assessment of undergraduate learning that has been called for in recent years. In large measure, this inadequacy stems from the time scales involved in the accreditation process; typically, American colleges and universities are reviewed for accreditation renewal once every 10 years. Clearly, this time span prevents students, parents, and other stakeholders from using accreditation reports as effective means of understanding the current state of undergraduate learning at a given institution. Additionally, while undergraduate learning is a major part of accreditation practices, the sheer amount of information gathered for a full-scale review means that the attention paid to undergraduate learning gains specifically is relatively small. An

additional problem is that, since accreditation reviews tend to be site-specific projects, the means with which they review undergraduate learning often lack consistency across contexts, making it difficult or impossible to compare one institution to another. For these reasons, Purdue's accreditation can be seen as a necessary but not sufficient mark of its overall health and the health of its undergraduate programs.

A Controversial Catalyst: the Administration of Mitch Daniels

The assessment effort at Purdue University must be understood as part of a broader set of changes, ones brought about by a new presidential administration. This change in leadership represents the most obvious and direct catalyst for the current assessment effort, though this effort would take years to develop and implement. Following a 5-year term as Purdue's president, the distinguished physicist France A. Córdova stepped down, in compliance with a longstanding Purdue policy that dictates that universities presidents relinquish their position after they reach the age of 65. Córdova ended her term as Purdue's president on July 15th of 2012. On June 21st, Purdue's Board of Trustees elected former Republican presidential candidate and then-Indiana governor Mitch Daniels to succeed Cordova.

As is perhaps to be expected, given that Daniels is a lifelong politician, his election as president by Purdue's Board of Trustees was highly controversial. There were a variety of sources for this controversy: his status as a conservative Republican, his history of cuts to higher education in his role as governor, his lack of academic credentials to suit the position, and the nature of his election by the Board of Trustees, many of whom he had appointed himself. First, the selection of Daniels attracted attention and criticism from some who felt that it was inappropriate for a partisan

politician to take on a role as the president of a public university. Public universities are typically seen as nonpartisan entities that are meant to serve all of a given state's constituencies. While university presidents cannot be expected to hold or voice no political opinions, the Board of Trustees invited criticism by selecting a national Republican politician of such prominence. Much of this criticism came from members of the broader Purdue community. Marilyn Haring, the former Dean of the College of Education at Purdue, was so unhappy with the selection of Daniels that she withdrew a $1 million gift she had made to the university. She was quoted in local media as saying, "The appointment of a politician to head Purdue University is an affront — an insult — to the academic enterprise" (Kingkade). Writing in the political journal *Jacobin Magazine*, two Purdue professors, Bill V. Mullen from American Studies and Tithi Bhattacharya from History, wrote that Daniels was "part of a national project to dismantle the already-shrinking public sector and subject the lives of working people to the vagaries of the market" ("What's the Matter with Indiana?").

Specifically acute were criticisms of actions Daniels undertook as governor that had direct impact on Purdue University and public education in Indiana generally. Particularly awkward was the fact that Daniels and the Indiana State Legislature had cut hundreds of millions of dollars from the state's support to public universities such as Purdue, causing many to wonder why a politician who had a direct hand in reducing an institution's funding would be rewarded with the presidency of that institution. Mullen and Bhattacharya identify these cuts as key elements in the rise of the cost of attendance for Purdue's undergraduates. "During Daniels' term of governor," they write, "student tuition at Purdue increased nearly 100 percent due to state funding cuts, and student debt

reached a record high of $26,000 per student." These cuts were seen as part of a broader antipathy to public education writ large during Daniels' time as governor, including the implementation of one of the largest private school voucher programs in the country. Such vouchers necessarily involve the transfer of funds from public institutions to private hands. Another point of controversy lay in the efforts of Daniels to restrict collective bargaining and union rights within Indiana. A longstanding center of manufacturing, an industry with traditionally high union participation, Indiana's adoption of so-called "Right to Work" legislation in January 2012 marked a major evolution of the state's economy. Daniels's endorsement of this legislation has been ascribed, in part, to efforts to prevent union protests of the 2012 National Football League Super Bowl in Indianapolis. Union issues are matters of sensitivity in American colleges and universities, as they have traditionally been unionized at higher rates than the economy writ large (Hibel "What Does the History of Faculty Unions Teach Us About Their Future?"). Many Purdue faculty members, for example, are members of the American Association of University Professors, a national faculty union and advocacy group for faculty.

Additionally, the contrast between Córdova and Daniels was stark in terms of academic qualifications. Córdova is an academic of truly impressive credentials. Holding a PhD in Physics from the California Institute of Technology, along with several honorary degrees, Cordova is an established expert in astrophysics, the youngest and first female chief scientist at the National Aeronautics and Space Administration (NASA), a member of a raft of national associations and academies, and the chair of the Board of Regents of the Smithsonian, America's national museum ("About Dr. Córdova"). Daniels, in contrast, holds a Bachelor of Arts in Public and International Affairs from Princeton

University and a law degree from Georgetown University (“Mitch E. Daniels, Jr. Biography”).

Of particular interest when considering issues of assessment, Daniels also courted controversy through the perception that he entered the position with a higher regard for some academic disciplines than for others. From the very beginning of his administration, Daniels made clear his affinity for certain programs and departments within Purdue—and his attendant lack of interest in others. In a campus-wide email announcing his selection by the Board of Trustees, Daniels was quoted as saying "No institution of any kind means more to Indiana today or tomorrow as Purdue University. It educates at the highest level the engineers, scientists, agricultural experts and information technologists on whom our state and national success disproportionately depend” (“Trustees Elect”). An incoming president being so specific and limited in his praise for particular university programs and departments sent a clear message that the Daniels administration would favor certain areas of study more than others. This message cuts against the basic purpose of assessment: if the administration came into power with a pre-existing set of expectations about the programs at the university that perform the best, it calls into question the good faith of the assessment proposal entirely.

Perceived Needs and the Foundations for Excellence Plan

The current effort to pilot and implement a comprehensive assessment of Purdue University undergraduate learning arose from a confluence of historical and political conditions. At the heart of these developments is a set of perceived problems identified by members of Purdue’s higher administration, in particular President Daniels. These problems include:

- A lack of consistency in Purdue undergraduate education, leading to less certainty about the value of a Purdue degree.
- A lack of accountability measures to ensure that students at Purdue are learning adequately and meeting university standards.
- An inability to make an evidence-based argument to students and parents that a Purdue education offers an exceptional "value." ("Message from the President About Tuition"; "A Message from the President About the Purdue-Gallup Index")

All of these problems are in keeping with those identified in current national assessment movement (see Chapter Two, "The Higher Education Assessment Movement").

To counter these problems, Daniels and his administration appointed a commission and instituted a large-scale initiative for change called the Foundations of Excellence (FOE) plan. Announced on August 25th, 2011, the FFE plan articulated goals that Purdue should strive to reach to meet the challenges of 21st century higher education. The twelve recommendations of the Foundations of Excellence plan are as follows:

- Active in vibrant and intellectually challenging community
- Respect for diverse views, backgrounds, and experiences
- Establishing a solid foundation for success
- Self-efficacy, confidence, and resilience
- Supporting intellectual and personal growth
- Learning in and out of the classroom
- Everyone belongs
- We are all accountable

- Productive leaders and citizens ("Foundations of Excellence")

These recommendations, in and of themselves, are somewhat vague and aphoristic. It's hard to think of a college or university, for example, that would not want to develop intellectual and personal growth, or learning both in and out of the classroom. But the goals are interesting nonetheless. For one, it's notable that while the FFE plan calls for a "vibrant and intellectually challenging community," there is no specific mention of research in this document, remarkable for a document of this type about a research-intensive university like Purdue. This is in keeping with the focus of the Daniels administration on undergraduate education as the central mission of the university, sometimes expressed as a "return to learning."

The second to last item, "We are all accountable," has perhaps the most direct relevance to the question of assessment and the CLA+ initiative. Throughout the assessment push at Purdue, the Daniels administration has cast the need for standardized metrics of student growth in these terms, as an issue of accountability for university educators and staff. A *Lafayette Journal & Courier* article that described the conflict between Daniels and the Purdue faculty senate over the CLA (see below) reported that "Daniels calls student growth assessments an accountability tool Purdue should have — and shouldn't fear" (Bangert). Similarly, Daniels told a reporter for *Inside Higher Education* that "showing [learning] is a matter of 'responsibility and necessity'" (Flaherty). The implicit moral argument is strongly in keeping with the crisis narrative in higher education, and is lent credibility in part by one of the most glaring issues facing college education today, the rapid rise in cost of attendance.

Identified Issue: Cost of Attendance

As mentioned in Chapter Two, rapidly rising tuition costs have played a large role in the national assessment movement in higher education of recent decades. Nationally, public four-year colleges and university tuitions for in-state students have grown an astounding 225% from 1984-1985 to 2014-2015 ("Trends in College Pricing 2014" 16). Worse, the trend for public four-year institutions has outpaced that of private four-year institutions, with the former having increased 3.25 times in that span and the latter 2.46 times. As our nation's public universities are specifically intended to provide a quality education to students who cannot typically afford a private university education, such increases directly undermine the public purpose of the system. Recently, the national trend has slowed, with growth in tuition from 2004-2005 to 2004-2015 at 3.5% for public four-year institutions, in contrast with growth of 4.0% for 1994-1995 to 2004-2005 and 4.4% for 1984-1985 to 1994-1995 ("Trends in College Pricing 2014" 16). This change is likely a result of changes to the overall economy, with the financial crises of the late 2000s forcing colleges and universities to slow their increases in tuition. However, because of the attendant decline in incomes and wealth nationally, including for college students and their parents, this slowdown in the growth of tuition has not been sufficient to prevent a massive increase in student debt loads, with student loan debts growing by an average of 6% per year in the four-year span from 2008 to 2012 ("Student Debt and the Class of 2012" 1). Coupled with an unemployment rate for recent college graduates as high as 7.7% in 2012, the moral and practical consequences are clear.

Purdue University has not been unaffected by these general trends. In the decade from 2003-2004 to 2013-2014, undergraduate tuition and fees for resident students rose

from $6,092 a year to $9,992 a year and from $18,700 to $28,794 a year for nonresident students ("Trends in Academic Year Tuition and Student Fees"). According to the Project on Student Debt, Indiana students writ large hold debt after graduation 64% of the time, with an average of $27,886 in debt per borrower ("Student Debt and the Class of 2012" 4). Purdue West Lafayette students specifically hold an average of $29,121 per debt holder, with 51% of students holding some debt after graduation ("Project on Student Debt: Indiana"). Student debt figures are known for having high variance, with averages frequently being poor descriptors of many real-world student outcomes. Still, with so many Purdue students graduating with that high of a debt burden, it's fair to argue that the cost of attending Purdue is undermining the Morrill Act's directive to "promote the liberal and practical education of the industrial classes."

In response, the Daniels administration has acted aggressively to reduce expenditures campus-wide. In his second Message to the Purdue community, Daniels announced a tuition freeze for 2013-2014 and 2014-2015. This policy was later extended by the Board of Trustees through the 2015-2016 school year. In his Message, Daniels specifically invoked the stagnating economy as a key reason for implementing the tuition freeze. "We must never forget," wrote Daniels, "that the dollars we are privileged to spend at our university come for the most part from either a student's family or a taxpayer" ("Message from the President about tuition"). Daniels specifically noted a few areas where Purdue could cut back, including redundancy, travel expenditures, and administrative overhead. In that last category specifically, Daniels announced that performance bonuses for senior administrators and professional staff with pay above $50,000 a year would be eliminated. This particular provision is noteworthy because

growth in administrative salaries is frequently cited as a key aspect of the rise of college costs. The ratio of college administrators to students stood at one administrator for every 84 students in 1974 but shrank to one administrator for every 68 students by 2005, and administrative costs grew by 36% from 1998 to 2008 (Ginsberg). By targeting administrative costs specifically, the Daniels administration acknowledged this source of university profligacy. His administration would go on to address a source of this administrative bloat: administrative redundancy.

Identified Issue: Administrative Redundancy

A typical reason for rising administrative costs, not only in university settings, lies in administrative redundancy. As a 2011 piece in *Inside Higher Ed* puts it, summarizing several major studies on administrative redundancy at large public university, "universities are complex, decentralized institutions. They waste a lot of money on redundant administrative activities and could probably save money in the long run if they made big changes to their structure" (Kiley). This redundancy is not hard to understand in the context of universities. Many are quite old, having had a century or more to accumulate programs, departments, offices, and divisions. They are often based on models of distributed leadership, separating control of curriculum from control of policies and fees, for example. Further, academic departments are typically allowed to work with a fair amount of autonomy, able to spend their allocated budgets as they see fit. All of these factors potentially contribute to multiple parts of the overall organization dedicated to solving the same problems. For example, Student Services departments at some universities might coexist alongside Undergraduate Life programs that substantially replicate the same work.

In March of 2013, the Daniels administration moved to address a perceived issue of administrative redundancy at Purdue. Specifically named were three areas: Undergraduate Academic Affairs, Student Affairs, and Housing and Food Services. The connection between these areas is obvious. Each relates to the day-to-day management of the undergraduate experience at Purdue, working to establish policies and procedures for more than 30,000 students as they attend classes, take advantage of extracurricular activities, and eat and sleep on campus. Specifically, the consolidation was enacted to combine six units: enrollment management, health and wellness, campus life, ideas and culture, learning, and academic success. These units were folded into the authority of the Provost's office. In a letter signed by both Daniels and then-Provost Tim Sands, this administrative change was explained, arguing that the consolidation would "align units that have similar missions, reduce confusion for students, effect more direct impact on student success, and emphasize programs that deliver innovative pedagogies" (Daniels & Sands). Academic and student services within the academic colleges and departments themselves were unaffected. Unmentioned in the letter, but an inevitable part of any effort to curtail administrative costs, were the inevitable cuts to employment that such a change would bring. I made several inquiries to higher administration to quantify the number of job losses or overall reduction in salary spending that resulted or will result from this consolidation; they went unanswered.

Identified Issue: A Campus Divided

A related issue to administrative redundancy lies in a peculiar aspect of Purdue's structure, procedures, and culture. Throughout my research, members of the Purdue community have identified a lack of consistency within Purdue's undergraduate

education as an impediment to the university's growth. One of the recurring themes of this research has been the modular, disconnected nature of Purdue's bureaucratic and institutional systems. As Brent Drake, Chief Data Officer in the Office of Institutional Assessment put it in an interview conducted on August 8 2014, Purdue "is more like 11 private colleges with an informal connection than one large public university." (Drake's interview is attached in Appendix C.) That is, each college within the university has traditionally had so much autonomy that the university writ large has lacked a consistent identity as a cohesive unit. As an example, Drake points out that Purdue did not have a provost until the 1970s, after more than 100 years of existence. As part of the role of a provost is typically to organize and administer curriculum and instruction, the lack of a provost was indicative of the lack of coordination and authority that has characterized the university's undergraduate programs. The disconnected nature of Purdue's administrative systems and institutional culture cropped up again and again in my research; many within the institution identify a lack of cohesiveness and interoperability between different colleges and majors, resulting in difficulties in communication and difficulty in navigating the bureaucracy of the institution for students and employees alike.

This divided nature manifested itself most powerfully in Purdue's complete lack of a core undergraduate curriculum for most of its history. Many universities have traditionally employed a set of guidelines and requirements that all students must complete, regardless of college or major. Such requirements typically include introductory-level general education classes, electives, and a minor or emphasis in addition to the requirements for a given major laid out by specific departments. While such curricular requirements are very common in American higher education, for the

large majority of its history Purdue's colleges and departments have had free rein to determine their own curriculum, with no central authority dictating equivalent standards. The consequences for assessment are clear: without consistency in instruction or educational expectations, there has been little basis through which to fairly and reliably evaluate the relative performance of various programs and departments. A key aspect of learning assessment lies in assuring that such assessments are fair and that they compare like with like. If different students within an institution are taking a significantly different curriculum, the lack of consistency hampers the tasks of judging how well they are learning and using that information to meaningfully direct pedagogical decisions. Even prior to the Daniels administration, an effort was afoot to consolidate and standardize key portions of the Purdue undergraduate curriculum.

An Early Reform: the Core Curriculum

One of the first major changes to Purdue's undergraduate programs was made official in February of 212. For the first time in its history, the college was to adopt a core curriculum, a set of classes and subjects that all undergraduate students were expected to take, to begin with the class of 2016. Although very common in American university education, a core curriculum had never before been implemented at Purdue, an artifact of the previously-mentioned lack of strong institutional standardization in undergraduate education. Prior to the implementation of the core curriculum for incoming freshman in Fall of 2013, different colleges, departments, and majors had complete freedom to institute whatever curricular standards they wanted. While this likely pleased some faculty members within those units, there were several negative consequences. First, because the actual educational experience for different Purdue students could vary so

dramatically by major, there was little ability to say with confidence what a Purdue education entailed. Second, a lack of commonality in requirements meant that students who switched majors were often significantly hampered in their attempts to graduate in a timely fashion. With the potential for little overlap in coursework from one major to the next, students could invest significant time and money in classes that they would later be unable to use towards graduation requirements if they transferred to another college at Purdue. In his interview, Drake noted that "the act of moving from Technology to Engineering here is like transferring to an entirely separate institution." This lack of transferability was a particularly vexing issue given that Purdue's First Year Engineering program has a very high attrition rate, leading many students to change majors in their third or fourth semesters. Additional problems were reflected in an 2012 report by the American Council of Trustees and Alumni, which assign Purdue a "D" grade for its curriculum "because of its lack of requiring adequate composition, literature, foreign language, U.S. History and economics courses for all majors" (Weisberg).

Luckily, a committee of faculty and administrators had already been at work on a Core Curriculum for over a year at the time this study was published. Chaired Dr. Teresa Taber Doughty, a professor of special education at Purdue, faculty representatives from all of Purdue's colleges worked to develop a set of provisional requirements to be implemented for new students beginning in Fall of 2013. In a proposal submitted to the University Senate Educational Policy Committee (EPC) which I acquired for this research, the Core Curriculum Committee argued that the "need exists at Purdue University to provide a means by which undergraduate students share a similar educational experience and in so doing achieve a set of common goals or outcomes

required of all graduates" ("Motion to Approve the Core Curriculum" 1). This Core Curriculum requires 30 credits worth of courses in various areas and disciplines, which are to be taken by students of all majors. Most of these credits can be satisfying from choosing from a limited number of courses that fulfill the given requirement. The Core Curriculum proposal went up for a vote before Purdue's faculty senate on February 20th, 2012, and approved. Students from the class of 2016 will be the first to complete the Core Curriculum as a graduation requirement.

While the Core Curriculum is not itself an assessment initiative, the connections to assessment are plain. First, one of the major difficulties of assessing collegiate learning is that students can take a vastly different curriculum depending on major. Not only are between-college variations potentially huge, even within-college variation can be quite large. In part, this curricular diversity is an artifact of the special training that students are meant to receive in their majors, especially in STEM programs. Instruments like the CLA+ attempt to assess general critical thinking skills as a means to avoid this problem, although the degree to which this is possible is disputed. A core curriculum like the one instituted by Purdue can help ameliorate this within-institution variation; although as the Core only makes up 30 credits, it will typically cover only about a quarter of an average students full credit load at graduation. Second, the Core Curriculum demonstrates an increasing amount of external influence on undergraduate learning within the university. The national college assessment movement frequently involves politicians, policymakers, and related stakeholders imposing standards on colleges and universities from above. The overall tendency is to move from more institutional and departmental autonomy to more and more control by outside forces, such as state Departments of Education that enact

curriuculum requirements. This additional influence can be spun positively or negatively, but there is little question that the assessment push both nationally and at Purdue specifically involves loosening faculty grip on actual educational standards and practices. In that sense, the Core Curriculum can be seen as part of a broader trend of central administration taking a more active role in shaping the collegiate education of the average Purdue undergraduate. As the Core Curriculum was being approved and implemented, the next stage of the Daniels administrations reforms—and the most controversial—was well underway.

The Initial Assessment Push

Though it would later grow into a major on-campus controversy that received national news attention, the assessment initiative that the Daniels administration spearheaded started out quietly. In spring of 2013, President Daniels appointed the Student Growth Task Force (SGTF) and charged it with finding a cost-effective way to accurately measure how much Purdue undergraduates were growing intellectually in their time at the institution. The 17-person committee included both academic faculty and administrative staff, including several experts in educational testing and assessment, including Drake and Diane Beaudoin, Director of Assessment in the Office of Institutional Assessment. The committee was co-chaired by Kirk Alter, the Vice Provost for Undergraduate Academic Affairs, and Jeff Karpicke, an Associate Professor of Psychological Sciences. Whitaker would come to be seen as the central figure leading the assessment push and, in an unofficial capacity, as Daniels's representative within the committee.

Daniels's specific instructions included in part:

> Over the past two years our faculty, led by the University Senate, has developed a remarkable core curriculum designed to help students achieve critical learning outcomes, outcomes highly valued by society and by employers – outcomes that will serve them well throughout their lives. I applaud the work done by the faculty. However, our future success requires that we clearly define this Purdue value equation and work every day to deliver that value.
>
> The faculty's work to date defines what we expect students to learn at Purdue and what we expect them to know and be able to do as Purdue graduates. It is not enough, however, to demonstrate that our graduates are society-ready. Important as it is to attract well-prepared students, how do we measure and demonstrate the value that Purdue adds? How do we measure a student's intellectual growth? How do we document that Purdue is continuously adding value to the learning of our students? ("Student Growth Task Force Memo")

Daniels went on to say that he was disinclined to set a hard deadline, knowing that the task could take considerable time, but asked for a "first iteration" by July 1st of 2013. He also wrote that he understood the process would be iterative and require on-going adjustment after initial implementation.

Several aspects of this missive stand out. First is the repeated emphasis on value as the core interest of assessment and the criterion of greatest interest to students, society, and employers. This appeal to "value" has been a commonplace in his public statements to and about Purdue University. Notions of value lie somewhat outside of traditional

notions of the purpose of higher education, which have often been defined in terms of liberal values that are more concerned with morality, ethics, and civic and political virtues such as an informed and engaged democratic citizenry. In other words, Daniels seems to embrace a vision of education that is concerned with *value* rather than with *values*. This shift in emphasis seemed to confirm the concerns of those who feared that his political conservatism and background in business would result in a corporatist educational philosophy. Mullen and Bhattacharya expressed this fear in writing that "universities are being remade, as Daniel Denvir pointed out, 'to operate according to the principles that guide multinational corporations'…. This means that we no longer 'teach students,' but 'provide a service to consumers'" (Mullen and Bhattacharya). As Mullen and Bhattacharya note, this type of thinking is particularly threatening to fields like the humanities and the arts, which frequently are oriented towards more abstract, less-material concerns than value in monetary terms. With his constant references to the value of a Purdue degree, Daniels lends some credence to fears like those of Mullen and Bhattacharya.

The SGTF developed three major requirements for any successful assessment system that would emerge from their committee. These requirements were:

> I. Expand the definition of "student success" to include not only completion of coursework but also overall "growth" experienced while at Purdue.
>
> II. Broaden attention from student inputs (especially skill levels of students admitted to Purdue) to a broader range of meaningful outcomes (including how students are prepared for productive lives at Purdue).

> III. Evaluate the ultimate success of Purdue graduates by more than whether or not they found employment. (*Understanding and Fostering Student Growth* 2)

These requirements were specifically expressed as necessary aspects of identifying effective "metrics," a term that reflects a social science and statistics background. In other words, the criteria by which the committee would define success were themselves cast in a particular methodological and epistemological framework, one that emerges from a more scientistic, more quantitative worldview. This worldview is unsurprising given the nature of the educational testing industry that dominates the world of higher education assessment and given the backgrounds of those on the committee, most of whom come from quantitative and STEM fields. Still, the emphasis on value instead of values reflects the inherent disadvantage that those of us in humanities fields like writing face; assessments of these types can generally be assumed to emerge from a basic philosophy of knowledge that is not shared by many of our scholars.

The committee set about investigating potential assessment mechanisms in short order. Although much of the work of the committee happened behind closed doors, interviews with committee members and affiliated staff involved in assessment at the university suggests that the task amounted to choosing a particular test or tests that had already been developed by outside entities. Choosing an existing test might seem an attractive option, given the inherent expense and difficulty at developing internal assessment mechanisms; but choosing a prepackaged test was not an assumed part of Daniels's initial directives. Nevertheless, the committee's primary activity involved

defining what types of growth should be measured, which instruments could assess those categories, and what the strengths and weaknesses of each identified instrument were.

Towards that end, the committee defined three "clusters" of student attributes that should be evaluated by Purdue's eventual assessment system: Personal Growth, Intellectual Growth, and Interpersonal Growth. Each of these clusters were then subdivided into multiple content areas, such as responsibility and ethical reasoning in Personal Growth, quantitative reasoning and critical thinking in Intellectual Growth, and written communication and teamwork in Interpersonal Growth. The SGTF's final report would later articulate their belief that this subdivision of clusters into smaller skills was necessary to choose appropriate assessment mechanisms. In this effort, they also solicited the input of the Educational Advisory Board, Gallup Education, and the Council for Aid to Education—the latter the developers of the CLA+.

The Roots of Conflict

Daniels's request for a July 1st, 2013 beginning to primary research proved optimistic. The SGTF would eventually announce the completion of a preliminary report in early October of 2013 and made an initial presentation to the Educational Policy Committee (EPC) of Purdue's faculty Senate on October 21, 2013. The EPC is the primary faculty body governing undergraduate education at Purdue. Its official role is defined as

> The Educational Policy Committee shall be concerned with, but not limited to: improvement of instruction, grades and grading, scholastic probation, dismissal for academic reasons and reinstatement, standards for admission, academic placement, the academic calendar, policies for

> scheduling classes, honors programs, general educational policy, general research policies, military training programs, general curriculum standards, coordination of campus and extension curricula, general academic organization, and interdepartmental and interinstitutional research and education programs. ("Standing Committees")

The EPC therefore is specifically designated by Purdue policy to control a broad swath of academic issues. Worth noting, however, is the absence of assessment listed among its designated responsibilities. With the growth of administrative structures such as the Office of Institutional Assessment, questions of responsibility and oversight will likely grow in the future.

Though the preliminary report would later be revised, the fundamental recommendations of the SGTF would remain substantially the same, and they would become the central object of later debate and controversy. The SGTF recommendations were multiple, in keeping with their division of skills to be assessed into clusters and subgroups. The basic outline of their preliminary proposal to the EPC is as follows:

- Assemble an Implementation Team, Evaluation Team, and Research Team, to supervise the collection of data, assure the accuracy of collected data, and explore the meaning of collected data, respectively
- Evaluate students on the Personal and Interpersonal Development clusters by "develop[ing] an index of non-academic factors related to student success"
- Assess the Intellectual Development cluster using the CLA+

- Research disciplinary competence in a manner developed or defined in-house by different disciplinary units and departments
- Develop an E-portfolio system that allows for student credentialing in various content areas using a digital "badge" system. ("SGTF Preliminary Report")

Several aspects of these recommendations are noteworthy. For one, though the final report would be far more specific in its recommended mechanisms, two of the three defined clusters are discussed simply through a recommendation to develop an index of non-academic factors. This recommendation is quite vague, given the specificity of the instrument identified as the appropriate tool for assessing Intellectual Development, the CLA+. Likewise, disciplinary competence, or the ability of students to demonstrate mastery of their major fields, is left to the hands of the various departments to handle themselves. This likely stems from a desire to reassure faculty that curriculum and evaluation of their majors will remain in their hands. Allowing departments to determine proficiency within majors also likely reflects a belief that traditional practices of grading and credentialing are already sufficient for the purpose of assuring disciplinary competence. For whatever the reason, the relative lack of specificity, other than the CLA+ recommendation, in the initial presentation made by the SGTF to the EPC demonstrates the degree to which the entire SGTF plan would come to be defined by the CLA+. As controversy would grow in ensuing months, the focus on the CLA+ as the real heart of the assessment effort—for good or bad—would become more and more clear.

According to later reports, the EPC's initial reaction to the draft report was one of significant concern. At particular issue was concern over the faculty's primary ownership of curriculum and learning, which would become a source of significant controversy in the ensuing debate over the assessment effort. On October 28th, the EPC drafted a resolution specifically responding to the SGTF's preliminary report, titled "Resolution on the Draft Report of the President's Task Force on Measuring Student Growth." The resolution, though written in the typically formal language of academic legislation, amounts to a strong rebuke to the SGTF. Specifically, the document repeatedly asserted the faculty's control of curriculum and teaching, including the assessment of student growth; questioned the validity of the CLA+ as a meaningful indicator of college learning; and stated that the appropriate assessment of disciplinary knowledge lies in course grades, which are more robust and valid than digital badging or discipline-specific assessment mechanisms. The EPC provided a copy of the resolution to Whitaker in anticipation of presenting the resolution to the faculty senate. Whitaker requested that the EPC delay submission of the resolution until the SGTF could reconvene and amend their report in light of these objections.

The major difference in the final SGTF report that was submitted to the University Senate in November 2013 was a change in nomenclature and definition of duties for one of the three recommended teams. The Evaluation Team was rebranded as an Oversight Team, and specifically mentioned that this team should be appointed by the University Senate, demonstrating acknowledgment on the part of the SGTF that faculty oversight was precisely the concern of the EPC. Additionally, the final SGTF report included a terse section named "WHAT THIS EXPLORATION IS NOT," which reads in

part, "The role of the Student Growth Task Force has been to [sic] limited to exploration and recommendation. The results of the work of the Student Growth Task Force should not be a change to the faculty province of curricula, learning and credentialing" ("Understanding and Fostering Student Growth" 3). This language clearly reflects efforts on the part of the SGTF to assuage the concerns of the EPC and faculty writ large concerning about the professoriate's control of undergraduate education.

The final report again endorsed the CLA+ as the primary means of assessing student intellectual growth. Additionally, the report spelled out which specific instruments could potentially be used to measure Personal and Interpersonal Development, although these recommendations remained more provisional than the continuing specific endorsement of the CLA+. Recommendations for Personal Development included using pre-generated scales such as the Basic Psychological Needs Scale, the Learning Climate Scale, the Self-Regulation Questionnaire, and the Gallup "Outcome Measurements" ("Understanding and Fostering Student Growth" 5). The exact means through which these scales would be implemented was not delineated in the report, but the committee recommended inviting Gallup Education to undertake this analysis. For Interpersonal Development, the committee specifically focused on what they called Inter-cultural Competence, and recommended the Miville-Guzman Universality-Diversity Scale, as well as an internally-developed portfolio system called Diversikey, which was created by Purdue's Diversity Resource Office. Finally, the report called for disciplinary assessment, but expressed no more specific requirement than that "each program work to develop or procure a method of assessing student growth towards disciplinary competence" ("Understanding and Fostering Student Growth" 6). The vague

nature of this recommendation likely stems from the SGTF's reluctance to appear to transgress faculty control of disciplinary education and assessment.

The report also admitted to a list of potential problems with the analysis of their proposed assessment system. These problems, it should be noted, are all methodological and epistemological in nature. In other words, none of the expressed caveats are similar to left-wing critiques like those of Mullen and Bhattacharya, or are related to the defense of the humanities and a commitment to the traditional values of higher education, as expressed by Haring. Rather, they are empirical problems that stem from the very quantitative, social sciences focus that are typical of the educational testing community. The specific qualifications and limitations mentioned by the SGTF include: selectivity bias, or the fact that students are not randomly assigned to different majors and thus perceived differences in student growth across majors may be the result of differences in population; maturation, or the chance that student growth might merely represent the expected intellectual development of age, rather than improvement from Purdue's education; sampling and attrition, or the negative effects on data from few students participating or many students dropping out of the study from freshman to senior year; scaling, or the potential that growth will not occur equally at different points of the student ability distribution; ceiling effects, or the potential for high-performing freshmen to have less room for growth, thus restricting their later senior scores and reducing perceived growth; feasibility of implementation, or the various practical and administrative difficulties that might prevent the collection and analysis of data; and student motivation to respond, or the potential for students to lack motivation to perform well on tests, artificially depressing scores ("Understanding and Fostering Student

Growth" 7-8). Interestingly, the issue of student motivation was listed last in the SGTF report, and yet as discussed in Chapter Three, this issue is the greatest impediment to validity and reliability in tests like the CLA+. Student motivation would later become a chief source of concern for the Office of Institutional Assessment.

The minor changes made to the final SGTF report between its preliminary presentation to the EPC in October and its final submission to the University Senate in November proved inadequate to allay faculty fears. Following submission, the EPC responded to a faculty request to enumerate principles about assessment and the principles of faculty control of curriculum. These principles, although they do not mention the SGTF explicitly, amount to a strongly worded response to the SGTF report. These principles are as follows (emphasis original in all cases):

1. The **primary responsibility** for establishing and assessing curricula, student learning outcomes, and student intellectual growth rests with the **Faculty** of Purdue University.
2. Assessment efforts must remain sensitive to the proper diversity and variety of programs and instructional contexts and objectives at Purdue, and shall avoid **unnecessary centralization**, standardization, or oversimplification of curricula or of assessment.
3. Assessment instruments **and their use** must be **credible** and appropriate, especially when widely disseminated and relied upon.
4. Assessment must be **fiscally responsible**, weighing the potential benefits of assessment with the time and money they require.

5. Purdue, as a leading University in the 21st century, should **remain committed to identifying and reporting** useful information about **its many contributions** to students' lives (its "value added"), in a variety of balanced, credible, and fiscally responsible ways. ("Some Principles for Institutional-level Assessment and SG Measurement")

Clearly, these principles function as a statement of faculty ownership over assessment and as an expression of skepticism towards some aspects of the SGTF plan. Whatever hope Daniels and the SGTF had that the assessment initiative would be implemented without faculty resistance must have been extinguished by this point.

Piloting

The SGTF's response to the University Senate's list of principles was one of the most significant developments of the assessment initiative: proposing a pilot study with which to compare various standardized tests of college learning and with which to gather preliminary data on Purdue's student body. As the later report of the SGTF Oversight Committee would point out, "These recommendations were made to administration. No recommendations were made to the EPC or the University Senate" ("Report of the Student Growth Task Force Committee"). This statement, while free of explicit complaint about this development, suggests faculty frustration at being left out of the process. An undated document title "Student Growth Task Force Pilot Program Recommendation" advocated for a pilot study to use both the CLA+ and the General Self-Efficacy Scale to "provide external validity to an aggregate measure of Purdue

undergraduate student body's intellectual growth" (1). In other words, the initial piloting effort was conceived of as a way to measure the validity of these instruments, to assess their appropriateness for measuring student learning outcomes, presumably in anticipation of expanding the sample to a larger, census-style approach. Towards that end, the document called for a stratified random sample, drawn from across Purdue's various demographic and academic groupings, of sufficient size to enable longitudinal study (as attrition effects ensure that participants will drop out between freshman and senior year) and adequate statistical rigor. The pilot proposal called for 10% of each college's undergraduate population to be represented in the study, for an ideal sample size of 2,922 (2). This recommendation would prove in time to be wildly optimistic.

The pilot proposal document also called again for the formation of an Oversight Committee, and this time explicitly named the University Senate as the body that should appoint the committee. It further notes that the purpose of the Oversight Committee to "oversee the pilot program" and "ensure that results are being used in an appropriate way" (2). In an email to the University Senate sent on April 1, 2014, SGTF co-chair Whittaker requested that the Oversight Committee be appointed by the end of April. He also again reassured the faculty senate that the results would not be used to evaluate the effectiveness of any individual faculty members or specific programs, and to "Reinforce the role of the University Senate for oversight and ownership of curricula and its improvement" ("Email from Dale Whittaker"). The University Senate would go on to appoint a nine-member, all-faculty Oversight Committee, including two members of the original SGTF. Dr. Kirk Alter of the Building Construction Management program would chair the committee, and go on to be a key voice of faculty resistance to the CLA+

initiative. The "Report of the Student Growth Task Force Oversight Committee," published in December of 2014, was an essential source in the preparation of this dissertation.

In addition to this primary pilot study, Purdue solicited a proposal from the Educational Testing Service (ETS) to develop an additional piloting program. This program was intended to provide further evidence for the validity of findings in the main pilot study, as well as to deepen overall understanding of the current level of ability of Purdue undergraduates. A proposal document I acquired in the course of my research, written by ETS personnel, indicates that "the data from the pilot will be used to ensure the validity, reliability and fairness of our next-generation student learning outcomes assessments" ("Proposal for Purdue University" 1). The piloting was to focus on three major areas: Critical Thinking, Written Communication, and Quantitative Literacy. Not coincidentally, these areas are closely related to skills and abilities tested by the CLA+. In an interview Brooke Robertshaw, a Data Analyst in the Office of Institutional Assessment, indicated that the purpose of ETS's piloting was to create additional validation evidence of the main piloting effort, in part to help assuage community fears about the assessment effort. However, the ETS pilot program was never completed. Robertshaw indicated that she was unaware of any specific rationale for canceling this additional piloting program, but she suggested that cost concerns, coupled with a conviction that the main piloting program was sufficient, were likely the reason.

Between the time of Alter's April email requesting the appointment of the Oversight Committee and the implementation of the pilot study at the beginning of the Fall of 2014, the specific format of that study changed markedly. During the summer of

2014, the pilot plan grew to include two other tests of general critical thinking alongside the CLA+: the Collegiate Assessment of Academic Proficiency (CAAP) developed by ACT corporation, the company that develops the ACT test for high school junior and seniors; and the California Critical Thinking Skills Test (CCTST), developed by Insight Assessment, a for-profit company that develops various tests of critical thinking and education. The plan also included the Global Perspective Inventory (GPI) and Miville-Guzman University-Diversity Scale (MGUDS) instruments previously named in the SGTF report.

The addition of these critical thinking tests represents a major change to the proposed pilot. After all, no tests other than the CLA+ were named in the SGTF report. The inclusion of these instruments was directed by the members of the Office of Institutional Assessment, who determined the actual research plan in the summer of 2014. Drake indicated that they specifically expanded the number of instruments considered in response to concerns from the Oversight Committee. "The Oversight Committee did not like that idea at all," said Drake in an interview, referring to the original plan to use the CLA+ exclusively for testing critical thinking. "They did not like the idea of putting all of our emphasis on one instrument. They believe—and I admit that I somewhat agree with them— the CLA+ in particular, we honestly don't know how well it works right now." The inclusion of these tests therefore functioned in part as a means to help assuage concerns of members of Purdue's faculty community. The tests could potentially cross-validate each other, demonstrating consistency in how they measure critical thinking and providing evidence that such tests provide meaningful information about a definable

construct. Additionally, giving faculty choices would potentially increase buy-in and help reassure them that they have an active role in the assessment process.

With the pilot study designed and ready for implementation, the first real steps towards generating real-world student data were ready to be taken. The CLA+ initiative's most important administrative leader, however, would not be present for the actual implementation of this policy. Vice Provost Dale Whittaker, who co-chaired the Student Growth Task Force Committee and was seen by many to be the Daniels administration's chief advocate for the assessment push, accepted a job as the Provost at the University of Central Florida, beginning August 1st, 2014. Several of those I spoke to privately about this project suggested that Whittaker's efforts on the assessment project were a key factor in his career advancement at Central Florida. Whittaker denied repeated requests to be interviewed for this research.

Initial Results

In August of 2014, during the week-long orientation prior to the semester that freshman undertake known as Boiler Gold Rush, the university implemented the pilot study of several tests of collegiate learning. Students participating in Boiler Gold Rush were emailed to solicit their participation, and were given a $5 Starbucks gift card as an incentive to participate. By testing first semester freshman before their first week of classes, the pilot enabled OIA researchers to undertake later longitudinal analysis, once that freshman class had reached their senior years. Information on the pilot tests, the number of students tested, and their results are represented in Table 1. This data is taken from both the "Fall 2014 CLA+ Mastery Results" report prepared by the CAE, and from

a report prepared by Robertshaw, who ran the piloting effort ("Fall 2014 CLA+ Mastery Results; "Student Growth Task Force Preliminary Report").

Test Name	**Number Tested (n)**	**Average Score**	**National Average**	**Price/student**
CLA+	128	1157/1600[1]	1039/1600[1]	$35
CCTST	87	80/100	76.5/100	$10.22
CAAP	74	65.2/80	59.4/80	$14.75

Table 1. Results of Fall 2014 Critical Thinking Pilot Study

Each test developer also provided distributional data, showing student performance across the score range. The CAE and Insight Assessments provided histograms of student performance, which are provided here in Figures 1 and 2, respectively. The ACT does not provide histograms in CAAP score reports, but does provide distributional data, with which I created the histogram of results in Figure 3. Note that the scoring range for the CAAP is 40-80.

Figure 2. CLA + Results. Fall 2014 CLA+ Mastery Results Institutional Report: Purdue University

[1] The CAE previously set the CLA scale from 400-1600, in order to make scores more comparable to SAT scores, against which they are regressed to account for ability effects. However, the CAE states that it does not technically employ a cap on CLA+ scores and does not report an upper bound to its scale. The exact practical implementation of an uncapped, norm-referenced scale is unclear. See "CLA+ Technical FAQs."

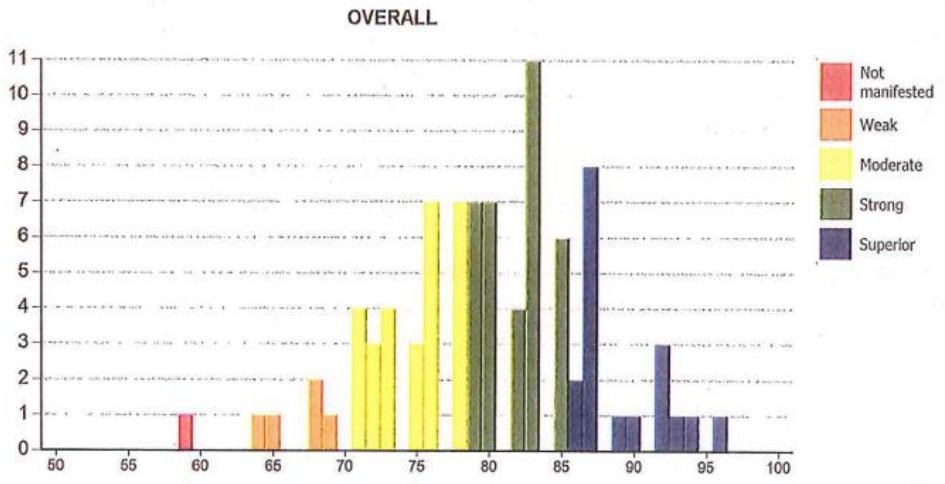

Figure 3.“CCTST Initial Results. California Critical Thinking Skills Test - Purdue University

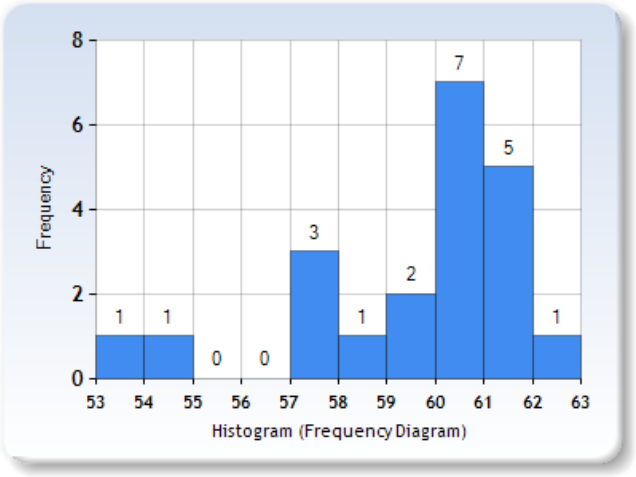

Figure 3. CAAP Initial Results. Collegiate Assessment of Academic Proficiency Test - Purdue University.

Immediately notable from this data is the far smaller number of students involved in the critical thinking pilot study, 289, compared to the number initially recommended by the OIA, 2,992. Less than one tenth as many students participated in the pilot study as

was called for in the initial SGTF report. This highlights again the difficulty in creating sufficient incentives for students to participate in these types of assessments. The enticement of a $5 gift card likely proved insufficient to attract students who were already dealing with the myriad responsibilities and time commitments of their freshman orientation. Whether this difficulty in attracting student participants should concern administrators is an open question. Certainly, given that Whitaker and others originally envisioned a census approach to the assessment effort, the difficulty in attracting students to participate in the pilot study should be of concern to administrators. Also notable is the cost of the various tests employed; at $35 a student, the CLA+ costs more than twice as much as any of the other tests utilized in the piloting program. As a public university that has committed to an effort to tighten its belt, Purdue has a natural interest in the cost of the standardized test it chooses.

It's difficult to draw meaningful conclusions from the available data. For each of the critical thinking tests, the students in the pilot sample were distributed fairly normally along the scoring ranges, with the average and median scores located somewhat higher than the national average. These results fit the intuitive expectation: with an annual acceptance rate that hovers around 60%, Purdue is a selective institution, meaning that the range of test takers in the pilot sample is restricted and negatively skewed. This limitation reflects part of the difficulty in undertaking standardized assessment of college learning: colleges and universities have vastly different student populations in terms of prerequisite ability. In fact, the entire undergraduate college admissions system exists precisely to create unequal student populations, with elite schools investing enormous time, effort, and resources in attracting only those students that are most likely to succeed.

To adjust for these differences in incoming population, the CAE regresses institutional CLA+ averages against SAT averages to compare to predicted scores based on incoming ability, and it determines a value-added measure by comparing freshmen scores to senior scores, as discussed at length in Chapter 3. However, the regression data the CLA+ calculates is not available for the current year, and as Purdue has opted to use a longitudinal design rather than the default cross-sectional design, there is no senior data against which to compare. Given that the CAE itself argues for the validity of its cross-sectional design, it is unclear why a longitudinal design has been an assumed aspect of the assessment since the project's inception. Given Daniels's repeatedly-stated regard for the book *Academically Adrift*, it is possible that this book's longitudinal design prompted this research choice. Regardless, with a longitudinal piloting design in place, it is clear that it will take several years before any meaningful data is acquired for this assessment project.

The CLA+ score report includes a survey on self-reported student motivation. These survey results are of interest, given that differing student motivation has been identified as perhaps the greatest potential challenge to the validity of such instruments. The CLA+ asked students "How much effort did you put into the written-response task/ selected-response questions?" For the Performance Task, 0% of students reported No Effort At All, 5% A Little Effort, 34% A Moderate Amount Of Effort, 37% A Lot Of Effort, and 24% My Best Effort; for the Selected Response section, 3% of students reported No Effort At All, 9% A Little Effort, 49% A Moderate Amount Of Effort, 26% A Lot Of Effort, and 13% My Best Effort ("Fall 2014 CLA+ Mastery Results Institutional Report: Purdue University"). It should be noted that self-reported data on

motivation is often considered imperfect from a reliability standpoint. For example, a 2009 review of this type of self-reported motivation scale found that "common measurement problems include a lack of representative normative samples, an absence of testing in authentic situations, and cross-cultural challenges due to differences in the definition and conceptualization of motivation" (Fulmer and Frijters 226). Still, this data at least provides a suggestion as to the pilot study participants' motivation on the CLA+ instrument. Given ETS's finding of "a substantial performance gap… between students in different motivational conditions" (Liu, Bridgeman, and Adler 352), it's difficult to say with confidence that the 40% of students who devoted less than A Lot of Effort on the Performance Task and the 61% of students who devoted less than A Lot of Effort on the Selected Response questions actually represented the best of their ability on the CLA+ pilot exam. Further, given the test-retest mechanism of the CLA+, whether students will engage with similar amounts of effort in their senior year testing remains a vital and unanswered question.

Robertshaw's report made limited claims about the pilot study. "The purpose of this work," wrote Robertshaw, "was to examine and compare three tests of critical thinking… and their ability to measure change in incoming Purdue freshman's to think critically" ("Student Growth Task Force Preliminary Report" 2). At the time of her initial report to the Oversight Committee, no results were yet available from the CLA+, limiting her ability to make claims about the viability of its use at Purdue. In terms of the tests of critical thinking, her report stated only that the CAAP and the CCTST would be capable of showing growth. In my interview and subsequent conversations with her, she suggested that this statement primarily meant that there was room for growth on the given

instruments. She later further suggested that the distributions of student scores along the range also provided evidence that the tests were functioning as effective instruments. Her statements to the Oversight Committee that the tests could be used to student growth were not intended to be a comprehensive claim about the validity, reliability, or practical viability of their implementation at the university.

Ultimately, the Fall 2014 pilot study amounted to an effective "dry run" of tests of critical thinking at Purdue, but provided little in the way of new information to those involved with choosing and implementing a standardized test at the university. That Purdue's incoming freshman perform fairly well on standardized tests of critical thinking should not be surprising, given Purdue's standing as a fairly selective institution. Without senior data against which to compare freshmen results, few conclusions can yet be drawn about the "value added" of a Purdue education, at least in terms of critical thinking. Further, student growth will have to be compared to national averages and averages of comparable institutions for results to have practical, meaningful value for various stakeholders. Given that it will be three years before a majority of the students who participated in the pilot study are ready to take the senior administration, the long-term nature of an assessment project such as this becomes clear.

Internal Skepticism

By the time Whitaker left for his new post, the large-scale implementation of the CLA+ at Purdue seemed assured. Although many questions were left to be answered, particularly concerning specific logistical issues concerning the implementation of the test, almost all of the stakeholders I interviewed and discussed the CLA+ proposal with at Purdue spoke as if the choice of the instrument was a done deal. Certainly, Whitaker

must have left campus with confidence that his choice would be approved. For this reason, I was surprised to find that members of the Office of Institutional Assessment expressed skepticism about the viability of these instruments for truly measuring student growth. Each spoke not from a specific institutional capacity as spokespeople for the OIA, but rather as individuals who are experts in the field of educational assessment. During our interviews, their attempts to speak with candor but with care demonstrated the difficulty of their positions as both researchers attempting to make responsible empirical claims and the fact that they operate under the directives of the Office of the President.

Drake, who stands higher on the administrative hierarchy than Beaudoin and Robertshaw, was most amenable to the use of these instruments as measures of college learning. He did repeatedly stress, however, that these instruments are still experimental and potentially misleading. He also spoke about the dangers of "institutional momentum," the potential for an instrument like the CLA+ to become the default choice at Purdue simply because it has been talked about the most, had the most resources devoted to it, and occupied the attention of the most people. Simply choosing the most talked-about exam could potentially lead, in Drake's view, to a kind of tunnel vision where other potential instruments are not given adequate consideration. He also admitted that student motivation, identified in the research literature as a major challenge to the validity of these instruments, was a potential problem. Speaking prior to the pilot study, Drake expressed concern and skepticism about being able to attract an adequate sample size and to provide appropriate motivation for students to give their best effort. "I don't think we're going to be able to do it real well…. [The provided incentive] in no way guarantees that the students will take it seriously, will put the effort in." Given the relatively small

sample size employed in the pilot, and the number of students who indicated lower levels of motivation on the survey, Drake's fears appear to have been well-founded.

Beaudoin and Robertshaw were both more direct in questioning the value of these types of tests. Beaudoin in particular expressed deep skepticism about whether any of these testing instruments could effectively measure college student learning, due in large measure to the lack of intrinsic incentive for students to perform to their best ability. As she said in our interview,

> "To be honest, I don't believe in the value added approach at all. I've watched students take standardized tests, three years ago as part of our Voluntary System of Accountability. I had to get 200 freshmen and 200 seniors to take the Proficiency Profile. I invited students, and had to do it in a proctored computer lab. I would proctor and just watch. I could sit and look in that room and tell you which ones were freshmen and which ones were seniors. Freshmen took the exam, they typically took between 45 and 60 minutes for a 60 minute test. You could see them scribbling notes on scratch paper. Seniors were typically done in 15 minutes. Click, click, click, done-- "can I have my $5?" Done. You're not going to see the value added because there's no incentive for the students to score well, take it seriously, so whatever scores you get.... I don't think they show value added."

As an alternative, Beaudoin argued that a more valid, more effective means of evaluating college performance is to look at outcomes several years after graduation. She argued that surveys of past graduates that ask questions about employment and financial success, life

satisfaction, and similar measures of well-being, taken at defined intervals such as 5, 10, and 25 years after graduation, would be a more authentic and more useful means to assess the value of a particular college education. Such results could be compared to other institutions and national averages to provide better information to students, parents, and the public at large about the value of individual schools. Robertshaw, for her part, stressed that there was a divide between her own epistemological orientation and the dictates of her current job. "I don't think you can measure critical thinking on a test," she said. However, she also said that her role required her to choose the best possible instrument, whatever her reservations, and that her goal was to do so to the best of her ability. Interestingly, she also confessed that the CLA+ is not her favorite of the three critical thinking instruments piloted, but declined to specify.

In one sense, the amount of skepticism expressed by senior administrators in the university's office specifically devoted to educational assessment is not surprising. Each of these researchers are experts with significant training and experience in the field of educational assessment and testing, and given the considerable controversy and criticism that these types of tests have attracted, their skepticism is to be expected. Given the potential stakes and resources involved in such testing, administrators expressing skeptical attitudes demonstrates a healthy desire for care in this type of initiative. On the other hand, given that the CLA+ initiative is ongoing, with the full, vocal support of the President's office, the amount of skepticism emanating from the Office of Institutional Assessment is surprising. Given that the very office tasked with implementing the assessment program at the university is staffed by researchers with profound reservations about the proposed tests, the continuing focus of senior administrators on critical thinking

testing demands scrutiny. Perhaps Robertshaw put it most bluntly and honestly in saying, "When push comes to shove, we have to do this…. When it comes down to it, if Mitch tells us we have to do stuff, [our concerns] have to get set aside."

Faculty Resistance

These notes of skepticism eventually grew into deeper resistance, this time coming from the faculty. Previously simmering tensions over the assessment initiative rose to a boil in Fall of 2014, largely prompted by a specific request from the Academic Affairs Committee of the Board of Trustees. On October 17th of 2014, that committee made a formal request of Patricia Hart, Chairperson of the University Senate, that the faculty choose one of the three critical thinking instruments "to be broadly administered by Purdue beginning in Fall of 2015, and then continuously thereafter" ("Report of the Student Task Force Growth Oversight Committee" 4). This action by the Board of Trustees functioned as a clear signal that upper administration intended to push forward with the assessment initiative, regardless of faculty calls for more time and more caution.

On December 19th of 2014, the Oversight Committee delivered its report to the Board of Trustees Academic Affairs Committee. That report strongly opposed the efforts of the Board of Trustees to speed the process. Speaking for the committee, chairperson of the committee Alter discussed the history of the effort at the university and identified the reasons for faculty resistance. The fundamental perspective of the Oversight Committee was summarized in Alter's PowerPoint slides: "Purdue should continue in a pilot/experimental mode rather than a broad and continuous implementation mode" ("Report to the Academic Affairs Committee" 8). This opinion stands in clear and direct conflict with the timetable proposed by the Board of Trustees, and outlines the most

direct source of conflict between faculty and administration. To justify this request for more time, the Oversight Committee:

- At this time no research plan exists, the experimental design has not been clearly articulated or vetted….
- A complete experimental research plan must be developed for this work to have validity.
- No faculty Implementation Team or Research Team has been assembled as recommended in the SGTF Pilot Recommendation.
- We do not yet have results of more than 1/3 of the cognitive tests administered - the CLA+ tests. ("Report to the Academic Affairs Committee" 8).

The Oversight Committee's full report went further, saying that fundamentally, "the Oversight Committee itself does not have the authority to endorse, on behalf of the faculty, any particular instrument for 'broad and continuous' use" ("Report of the Student Task Force Growth Oversight Committee" 6). Further, the report called the evidence assembled in the pilot study "thin," and argued that the Committee did "not have sufficient confidence to endorse any one of the critical thinking instruments for broad and continuous use" ("Report of the Student Task Force Growth Oversight Committee" 6). For their part, the Board of Trustees pushed back forcefully against the Oversight Committee's recommendations, later reported to have been "not buying it" and to have "picked apart" Alter and his co-chair, Patrick Kain.

This growing conflict remained a quiet, intra-institutional facet of Purdue life through 2014, but would soon grow to become very public. On January 27th, 2015, the local newspaper for the greater Lafayette area, *The Journal & Courier*, published a piece on the growing rift between the Daniels administration and the faculty senate titled "Daniels, Purdue faculty in test of wills." The piece concerned the growing faculty resistance to the assessment project generally and the CLA+ specifically. "[T]here's no question," wrote *Journal & Courier* reporter Dave Bangert, "that lines are being drawn between a Purdue administration that wants an annual measure of students' intellectual growth in place by this fall and faculty members who say they need another year to come up with a solid, academically valid standard" (Bangert). The piece went on to detail the general perspectives of the parties involved, with the Daniels administration calling for evidence to demonstrate the value of a Purdue education and the faculty expressing concern that these measures may lack validity and reliability. Daniels was quoted as saying that "this should not become one of those paralysis-by-analysis, permanent procrastination exercises," demonstrating frustration with the timeline proposed by the faculty, and arguing that faculty "get nervous about things they shouldn't be" (Bangert).

Faculty leaders pushed back. Hart argued that there was little reason to rush the assessment and that doing so would potentially undermine the findings of the assessment. She was quoted as saying "I heard the trustees say something to the effect of, 'Don't let perfect get in the way of good,'…. My response was, 'Don't let unacceptable get in the way of good'" (Bangert). She went on to point out that Purdue already undertakes a number of assessment efforts and gathers a great deal of data about student success already. The *Journal & Courier* story detailed Alter's argument to the Board that, if the

assessment process was slowed down and the best system of assessment developed, Purdue could earn national acclaim, analogizing this sort of research to the famous Framingham Heart Study.

There is a sense in which the *Journal & Courier* article became a self-fulfilling prophecy. By describing the conflict as a "test of wills," it's likely that both sides involved felt compelled to defend their "turf." The perception that this was a major conflict over the future direction of the university likely inspired deeper feelings of animosity, causing both sides to dig in their heels. Speaking under the condition of anonymity, a senior Purdue administrator who works in the broad domain of undergraduate education said in an email, "Look, Mitch is a politician. He's keenly aware of public perception. I don't have any insider knowledge but it's my assumption that the J&C article likely left him feeling backed into a corner." The very public nature of the conflict, following the *Journal & Courier* article, was exacerbated by the publication of a piece in the national industry website *Inside Higher Education*. The piece, published a day after the *Journal & Courier* article, largely echoed that piece in detailing the general history of the assessment push at Purdue and outlining the basics of the divide. While the piece added little new insight into the conflict, its presence in a major national trade publication that is followed by many within higher education demonstrates the keen interest that questions of assessment and control of curriculum attract.

For all of this attention and perceived animosity between the two groups, the explicit conflict between the two sides, as expressed in official policy documents, is a matter of timeline and not of principles. The Oversight Committee recommended a Fall 2016 launch of wide-scale implementation of an assessment mechanism, while the Board

of Trustees and Daniels requested a Fall 2015 launch. While Daniels was quoted by *Inside Higher Ed* as saying "I didn't hear from anybody who feels we shouldn't be accountable and shouldn't be taking any such measurements. I didn't hear that. I heard discussion about the best ways of doing this… we'll continue talking" (Flaherty). Additionally, all sides seem to agree that Purdue's undergraduates likely learn a great deal and believe that the outcome of any assessment effort will be positive.

The deeper question is the experimental nature of ongoing efforts. The University Senate has argued in terms of continuing to investigate possible solutions, taking care to ensure that results are valid and reliable and produce meaningful data about Purdue students. Hart was quoted as saying, "You have to have a very careful design that proves what you say it proves…. So this is quite different than a public opinion poll, a consumer poll or a poll about elections. This is research that will stand the test of time and stand up to scrutiny" (Bangert). In contrast, the pursuit of speed by the Daniels administration suggests a greater desire to simply implement some sort of system, with less concern for the actual value of the data created. In this perceived desire to assess first and ask questions later, Daniels recreated controversy from his term as governor of Indiana, in which he was a famously assertive champion of standardized testing in K-12 education. The CLA+ initiative, according to the *Journal & Courier*, reflects "Daniels' affinity for metrics and being able to boil things down into something more than hazy assurances of accomplishment. Before coming to Purdue, as a two-term governor, he'd championed similar, easy-to-read grades for K-12 schools in Indiana and tying teacher pay in part to student performance" (Bangert). This extension of typical school reform principles into the higher education sphere is precisely what early critics of the appointment of Daniels

to the Purdue presidency were afraid of. Given the previously-mentioned institutional inertia, meanwhile, the initial test chosen could quickly become a norm that would be difficult to set aside even if better alternatives were discovered, given that any new test would have to be compared to older results from different instruments. This concern invites the question of whether there has ever really been a meaningful opportunity to choose a different test than the CLA+.

Was the CLA+ Preordained?

A question of particular importance and sensitivity to this project lies in the selection of the CLA+ as the primary mechanism of assessment and whether alternative methods were ever seriously considered at all. Part of the difficulty of a project such as this lies in investigating ideas that are part of the ambient discussion but which are not formally expressed in public. In many casual conversations and off-the-record discussions, members of the campus community spoke straightforwardly as if the CLA+ was chosen prior to the piloting effort that was ostensibly intended to find the best instrument. That the CLA+ was always targeted as the tool of choice by the Daniels administration, in other words, has been an "open secret" on campus. None of my interview subjects and none of the official documentation I've found stated this directly, but the timeline and presence of the CLA+ on early documentation strongly suggests this to be the case. Beaudoin stated in her interview that "there was an initial sense that we were just going to do CLA+ and go with that," but did not state that the test's selection was preordained. In contrast, Brooke Robertshaw said regarding the purpose of the pilot study, "My understanding—and I'm the lowest person on this totem pole—my understanding is that we were looking to find out which test we wanted to use."

The previously-mentioned anonymous administration official wrote, "I think the writing was on the wall very early on that Mitch wanted the CLA. He's never made a secret of being a big fan of AA [*Academically Adrift*] and I think he saw himself as part of a lineage, part of a movement." Daniels's prior discussions of *Academically Adrift* lend credence to this point of view, as he has gone so far as to state in public the Arum and Roksa's book is his "bible." The *Journal & Courier* article about faculty resistance to the CLA+ initiative notes that "*Academically Adrift* fit squarely within Daniels' affinity" for standardized testing (Bangert). Given that the assessment initiative was a directive of the Daniels administration and the swift way in which the test was chosen, it seems reasonable to conclude that the SGTF were at least initially predisposed to select the CLA+ as its primary instrument for measuring intellectual development. It is worth noting, however, that Daniels told the *Journal & Courier* reporter that "I'm indifferent to what measuring tool we use or how we use it. That's an absolute classic question for the faculty to decide" (Bangert).

The possibility that the selection of the CLA+ was inevitable raises uncomfortable questions about the appropriateness of such preselection, particularly for a public university. The purpose of the piloting effort, as suggested by Robertshaw and Drake, and as specifically directed by the Oversight Committee, was to determine which test might be best. If an official committee of the University Senate dictated an open competition between different tests, but the choice was constrained from the start, it would suggest a lack of good faith on the part of the administration. Still, the Board of Trustees did specifically empower and request the Oversight Committee to choose one of the three tested instruments. The larger question is why the three tests were chosen. As mentioned

previously, the exact selection criteria for these three instruments were never explicitly detailed by the OIA. The Oversight Committee report complained about this lack of information, saying, "No information was provided regarding instrument selection criteria, experimental design, instrument design, actual instrument questions, or access to results and interpretation" ("Report of the Student Task Force Growth Oversight Committee" 4). In discussions with OIA staff, the impression I was given was that the three critical thinking tests were chosen because there are a limited number of possibilities currently commercially available, although it is not clear why some major competitors like ETS's Proficiency Profile were not included in the piloting effort. The somewhat ad hoc nature of the instrument selection process, along with the threat of institutional inertia, demonstrates the capacity for high stakes decisions to be made without a clear institutional justification.

Though the Oversight Committee never made a formal recommendation about which test instrument to choose, the continuing debate on campus would focus almost exclusively on the CLA+, demonstrating the power of initial impressions. The CLA+ was the topic of conversation at a Purdue faculty senate meeting where Daniels would again make his case.

Buying Time

At a packed University Senate meeting on February 1st, where the entire backroom gallery for non-members was filled with interested parties, both the University Senate and President Daniels were given another opportunity to make their case. In opening remarks, Senate Chairperson Hart argued again for more time and more vetting, arguing that "This is not paralysis by analysis. This is taking the time to get it right." Hart

specifically mentioned attrition, motivation, and basis for comparison as specific concerns about the CLA+ and the assessment initiative generally. Hart counseled Daniels to appoint a blue ribbon panel, jointly commissioned by the Provost's office and the University Senate, to oversee the creation of an internal longitudinal assessment instrument. Hart asked, "Mr. President, can you commit to giving us the necessary resources and autonomy to move forward, together?"

Daniels's response was a reiteration of his previous positions. "I want to move forward towards the goal of successful fruition of a goal that, as far as I know, we all agree on," said Daniels. "That's been the position of several committees within the faculty. That goal is to be an accountable university." He demonstrated frustration at the continuing delays in implementation of assessment. "This has been a long process," he said. He repeatedly stressed that he had no interest in removing faculty control of curriculum, and that the purpose of the assessment was not to evaluate the progress of individual departments or majors within the university. He again cited arguments like that of *Academically Adrift* that state that limited learning is occurring on college campuses, and argued that his only intent was to demonstrate that a Purdue education was a high value proposition. He then stressed what has become the real crux of disagreement, more than the timeline issue which is really a proxy for this deeper disagreement: he flatly rejected the idea of a proprietary assessment system developed internally at the university. "We need documentary evidence…. Very importantly, we need to be able to compare results with others. Therefore, producing some sui generis, Purdue-only exam wouldn't meet this criteria…. We'd have no ability to compare to anyone else." In this rejection of a system developed internally, Daniels echoed the reporting of *Inside Higher Ed*, which

read “Daniels said an internal tool “won’t fly,” since it’s important to be able to compare Purdue to other institutions” (Flaherty).

Ultimately, Daniels extended the timeline until the next faculty senate meeting, scheduled for April 20th, “because this is complicated business,” in his words. This extension of less than three months was far shorter than the faculty’s initial request of one year. “That should be plenty of time,” argued Daniels, “to make a suggestion, even if it’s something other than the CLA+.” He then pointed out that a true longitudinal test from freshman to senior year would take until 2019 to be completed, demonstrating the downside to taking more time to decide. “We’ve spent two long years of hard work on this. Let’s take that first step this fall.” Faculty members had several questions for Daniels. One question was whether the assessment mechanism developed at the West Lafayette campus would be ported to the other Purdue system campuses. Daniels was noncommittal, saying that it seemed to be a good idea but that he “wants to respect your campus’s autonomy.” Another faculty member asked Daniels if his most important priority was assessing for internal purposes or for comparison to other institutions. Daniels again stressed that the intent was not to make comparisons between different majors and that the limited sample sizes would make this difficult. He did add, however, that “we’ve got some very clever people here and they may be able to make that happen.” This response was perhaps off-message, as the lack of between-major comparisons has been a point repeatedly voiced by the administration, likely as an attempt to forestall faculty concerns. “We need to add something discipline-specific, something portfolio based…. That’s what I’m talking about when I say something will evolve.” A faculty member asked specifically about the issue of student motivation, to which Daniels replied,

"It's a problem with at least 100 different solutions, though it is a problem…. Somewhere between free pizza and a thousand dollar bill, there's gotta be a solution [to motivating students]." This response perhaps demonstrates a lack of understanding of the depth of the motivation problem, given the focus on that problem in the research literature. After extensive questioning, the faculty senate moved on to other business, and the assessment effort remained in limbo until April.

On March 5th, the local public radio station WBAA ran a story on the current state of the assessment conflict, for which I was interviewed. The story raised the possibility of Purdue becoming a model for other institutions, and thus spreading the national momentum for higher education assessment. "It seems clear that schools across the country are adopting similar tests all the time. And recruiter Roger Malatesta says companies might soon follow suit for the same reason schools like Purdue are considering the test – because many of their peers are already on board" (Jastrzebski). This continued movement toward large-scale assessment testing highlights again the national focus on Purdue's assessment efforts, and the high stakes involved in these types of decisions. My own comments in the story indicated concern over the problems of motivation and attrition, which are typically raised in regards to these types of assessment. In an email to me following up on the WBAA story—the only time Daniels commented to me for my dissertation—he wrote,

> Don't know anyone who thinks this is easy or without challenges. But the science is advancing, and at least on the basic question of critical thinking it cannot be beyond our collective capacity to get a reading. The same

> faculty panel that recommended CLA+ agreed with you on the need for augmentation with discipline-specific measures.
>
> …I have seen years of resistance to accountability by any method in the K-12 world. After "It's too difficult", the next trench retreated to is "We'll make up our own test", which then is useless because it provides no basis for comparison.

Again, the continuity with K-12 reform is clear, and given the controversy and attention that such testing has engendered, the political stakes are clear. Whether Daniels will have the transformational impact on higher education that he has had on Indiana's other public institutions remains to be seen.

The Road Ahead

As I completed this dissertation, an unexpected development occurred. On April 1st, the Oversight Committee submitted a formal proposal to the Daniels administration, requesting a one-year postponement on full-scale critical thinking testing. Rather than rolling out the full assessment system this fall, the Office of Institutional Assessment will instead attempt a pilot of some 360 freshman, using a commercial test of student learning such as the CLA+. Meanwhile, campus stakeholders will work to build a consensus definition of critical thinking to guide future testing efforts. In a new article for the *Journal & Courier*, Alter is quoted as saying "This is a good compromise between the parties. The president and Board of Trustees get the next phase of standardized testing, ….and the faculty gets the assurance that we will pursue this from a much more thorough and academically sound approach" (Paul).

Given that the timeline was the main source of contention between the faculty and the Daniels administration, at least ostensibly, the faculty appear to have won the short term intra-institutional conflict. Slowing down permits faculty members and committees to develop metrics that they are more confident in. Meanwhile, Daniels and his administration get a renewed commitment to implementing a large-scale assessment of the kind that they have long argued for. In the broader view, however, key questions remain. As indicated previously, the issue of timeline was often discussed in my conversations with faculty as a proxy for larger issues of best practices in assessment and faculty control of curriculum. While this latest decision has bought all parties some additional time, it is likely that the deeper concerns will persist, not only at Purdue, but in the American higher education movement writ large.

The history detailed herein demonstrates the tangled, contested ways in which national educational movements like the higher education assessment push are actually implemented in real-world local contexts. Rather than being a smooth progression from ideas debated on the national stage to specific, actionable policies that are executed in a straightforward manner, educational reform involves constant negotiations, small and large, between various stakeholders involved in the world of college education. These tensions and conflicts—educational, institutional, empirical, philosophical, theoretical, political—have profound consequences for writing studies specifically and higher education generally. Those consequences, and what we as writing educators should do in response to them, are the subjects of this dissertation's final chapter.

CHAPTER SIX: CONCLUSIONS

For the past eighteen months, I have researched the higher education movement, the CLA+, its relationship to writing studies and the educational testing industry, and the implementation of the test here at Purdue. From the outset, my desire has been to investigate all of these in the spirit of balance and fairness. I cannot claim objectivity on the broad education reform movement, the political forces that have agitated for assessment of colleges and universities, or the administration of Mitch Daniels. But I have attempted to remain open-minded about the political and policy initiatives my research concerns, and to weigh the various pros and cons of both Purdue's assessment initiative specifically and the broader higher education assessment movement generally. This chapter details my own analysis on these and related topics after the past year and a half of research and consideration.

Some Form of Assessment is Likely Inevitable

One of the most obvious conclusions that can be drawn from this research is that, both at Purdue and in the United States higher education system writ large, some form of student learning assessment is likely inevitable. The forces behind this movement are powerful and unlikely to be completely stymied in their efforts. On the national level, successive presidential administrations have made assessment of college learning a national issue, and have articulated high-stakes accountability systems that would be hard

for any college to ignore. Even the most deep-pocketed private universities could ill afford to ignore the Obama administration's efforts to tie access to federal aid to college rankings that would depend in large measure on standardized tests. That particular proposal is a controversial initiative proposed by a controversial administration, and it is possible that these rankings will not come to fruition. But with such consistency between both Republican and Democratic administrations, and such widespread agitation for change at the top of our educational policy apparatus, there is little doubt that some forms of assessment are likely coming, most likely to public universities that are beholden to state legislatures and governors.

Similarly, at present there is little doubt that Purdue University will enact some sort of standardized assessment of student growth in the near future. After all, the University Senate has committed itself to taking part in the development of an assessment; their primary disagreement with the Daniels administration concerns the exact method of assessment, how the results will be interpreted and used to affect policy, and the ultimate control of assessment issues within the institution. For his part, Daniels has never wavered in his commitment to enacting some sort of standardized assessment system at the university. As a university president who sees himself as a transformative leader in the American university system—and as a former politician who has been widely rumored to seek a role in national politics in the future—Daniels likely sees the CLA+ initiative as a clear example of his commitment to meaningful reform. Given all of the "institutional momentum," to use Drake's phrase, I would be deeply surprised if Purdue was not enacting a persistent assessment project in Fall of 2017 at the latest.

What remains to be seen in both the national and local context is what form these assessments might take. As much as it may seem like change is inevitable, and as much as this change might frustrate those of us who think there are deeper goals in college learning than improving narrowly defined educational skills, my research indicates that compromise and influence are possible. Though I feel that the Daniels administration has agitated for a very specific set of consequences from the outset, I also believe that they are genuine in their invitations to the University Senate to influence the process. Similarly, policy leaders such as those in the Obama cabinet do make genuine efforts to include a variety of voices from within the community of college education. The question in both local and national contexts is how constrained potential options are. With standardized testing such a major aspect of contemporary educational policy, stakeholders might always gravitate towards those types of instruments despite the myriad issues with their use. The faculty at Purdue has been given a choice, but as members of the faculty have pointed out, that choice amounts to one of three tests of critical thinking, none of which has had long-term vetting or a great deal of external review. It would be understandable if some in the faculty thought that this represented no real choice at all.

But space still remains for a strong faculty role in the development and implementation of any assessment system at Purdue. All involved acknowledged that the process would be long-term and ongoing. Daniels, for his part, stressed that they would take a "learn by doing" approach to the assessment, suggesting that his office would take results with an appropriate amount of skepticism. That philosophy leads to another essential conclusion: interpretation of results is as important, or more important, than the

specific assessment system that generates those results. The qualifications and limitations included in both the CAE's own research and third party research demands care when interpreting results. As mentioned, the potential confounding factor of student motivation might deeply impact our confidence in the validity of the CLA+'s findings. Issues of attrition, scale, ceiling effects, and other potential problems in this type of assessment must be taken into account when the administration and other stakeholders review the outcomes of the test, if in fact the CLA+ is implemented at the university. Additionally, the CLA+'s criterion sampling approach suggests that results should not be interpreted to draw conclusions about the performance of different departments or majors, as the administration has pledged it will not do.

Ultimately, only ongoing collaboration between the Faculty Senate, the Office of Institutional Assessment, and the President's office can ensure that any assessment system is applied consistently and fairly. That will require a spirit of mutual trust and a willingness on all sides to compromise, at times. The need for compromise does not mean that the faculty should abandon their objections, and it does not mean that agreement on all aspects of the assessment system or its results will ever be achieved between faculty and administration. But if the President's office honors its commitments to keeping curriculum in the hands of faculty, then each side can potential serve as a useful check on the other. For this reason, the continued efforts of an Oversight Team, as called for in multiple proposals from the SGTF, are crucial to the long-term health of assessment at Purdue.

Critical Thinking Measures are Inadequate

One of the interesting facets of this research lies in the general focus on the CLA+ and other tests of critical thinking, despite the fact that the assessment program also includes calls for tests of intercultural knowledge and disciplinary education. Again and again, people involved with the assessment initiative on campus took the potential tests of critical thinking to be the central issue at hand, largely ignoring the disciplinary and intercultural aspects of the assessment. This focus extended from Daniels as expressed in interviews with the press and in emails, members of the Office of Institutional Assessment, faculty members, and assorted other members of the Purdue community with whom I consulted for this research. The news stories in the *Lafayette Journal & Courier*, Inside Higher Ed, and from WBAA all discussed only the critical thinking tests. Generally speaking, the crux of the assessment initiative and of the conflict between Daniels and the faculty senate was perceived to revolve around only one aspect of the plan proposed by the SGTF.

Why? For one, issues of controversy are issues that attract attention. Because the CLA+ and other tests of critical thinking are the aspects of the SGTF proposal that have elicited the most criticism, they are also the aspects that have drawn the most attention. This is particularly true when it comes to the press, whether local or national, as controversy and scandal are typically most likely to generate press coverage. Another reason for the focus on the critical thinking tests likely rests on Daniels's repeated invocation of *Academically Adrift* during his calls for accountability at Purdue. Because that text used the CLA as its primary method for assessing student learning, the focus on the use of its successor in the pilot study is to be expected, especially given the

controversy the book generated. Finally, the paucity of information about potential systems that might be used to assess disciplinary knowledge means that there is little to discuss in that regard. Because the SGTF repeatedly stressed that the assessment of disciplinary knowledge would be determined by individual departments or majors, there was not much content in that area to debate.

But if this focus on critical thinking is natural, it is also potentially dangerous. *Academically Adrift* caused a great stir, despite the many criticisms of its methodology. The book argued that the average college student showed little growth in his or her university career. In fact, the subtitle of the book is *Limited Learning on College Campuses*. This created a widespread impression that college students learned very little, calling into question public investment in higher education. But this popular impression ignored the fact that the CLA was specifically designed to elide disciplinary knowledge. In other words, the test was not designed to assess whether a history student learned disciplinary knowledge in history, a biology student in biology, and so on. As a result, the average person likely saw *Academically Adrift* as a more damning critique than it really was, as few likely understood the difference between critical thinking and disciplinary knowledge. An undue focus on any critical thinking test at Purdue could potentially result in an underestimation of the quality of a Purdue education.

In his negative review of *Academically Adrift*, Richard Haswell reflected on the problem with over-relying on a particular, individual test like the CLA:

> Like the refrain from a sea chantey, the phrase "critical thinking, complex reasoning, and writing" runs unvaried throughout the book. The phrase is unvaried because, amazingly, the authors fuse all three together into one

> lump. Their one and only measure of this lump is CLA's Performance Task, which gives students ninety minutes to read some documents and then write a genre specific response…. One number to represent "critical thinking, complex reasoning, and writing"! The methodology is antediluvian from the perspective of development and composition experts—none mentioned in AA—who have studied undergraduates' growth in critical and literacy skills and found it interactive, differential, and sometimes regressive. (489)

If assessment mechanisms like the CLA+ are inevitable at Purdue and elsewhere, then it behooves the faculty to create alternative, discipline-based assessments that can potentially corroborate and deepen positive findings or complicate negative ones. All assessment systems require validation. Disciplinary learning could be shown to be related to critical thinking metrics like that of the CLA+. Alternatively, we could learn that some students acquire a great deal of disciplinary knowledge that might help them professionally while not showing much growth in critical thinking. One way or another, for the long-term health of the faculty and of the American university systems, we must ensure that limited, reductive instruments do not become the sole way in which college learning is assessed. To do so risks too much, particularly in an economic and political context in which the reputation of the academy is already in question.

The CLA+: Could Be Worse

My attitude towards the CLA+ has been one of the enduring questions that has attended my dissertation research. Frequently, because of my political and educational disagreements with the education reform movement generally and the standardized

assessment push particularly, people I discuss this project with assume that I reject the test out of hand. This is not the case. I have a complex perspective on the test, one which depends a great deal on institutional and political realities that agitate for such testing. In general: though I find the test limited and reductive, and would prefer that it never be used in a high-stakes assessment, given that testing of this type is likely inevitable in many contexts I endorse it in comparison to many alternatives. In other words, if we must use a standardized instrument of critical thinking to assess college learning, I would prefer the CLA+ to some other possibilities.

In my view, the test has the following strengths:

1. The CAE is made up of "the good guys." Researchers like Richard Shavelson, Steve Klein, and Roger Benjamin have had long careers within the academy and have demonstrated a considerable personal and professional investment in the higher education system. In contrast, many standardized tests of education are developed by for-profit companies that are staffed primarily by members of the business community. While the CAE's non-profit status means little for its actual ability to accumulate profit, I firmly believe that the members of the institution are deeply committed to their project of improving college education.
2. The test involves real student writing. As this dissertation has documented, timed essay responses of the type that are utilized in the Performance Task are frequently challenged in the writing studies community, usually on validity and authenticity grounds. But essentially all scholars within writing studies would still prefer such essay tests to entirely indirect tests of writing, such as multiple-choice questions about grammar or editing, as are used in some educational tests. If tests

are coming, then we should advocate for tests with as much actual writing as possible. With writing fighting for its institutional legitimacy in an academy that has enacted steep cuts to the humanities, meanwhile, the importance of writing on these tests can help demonstrate the importance of our teaching.

3. The criterion sampling philosophy of the CLA+, if it is taken seriously by administrators, helps to protect against drawing conclusions about individual programs and majors based on the limited samples typically utilized in this type of research. As I've argued, the interpretation of these tests is as important as which test is chosen. Since the CAE's own documentation argues that collegiate learning is a holistic, multivariate phenomenon that cannot be attributed to specific majors or programs, the potential negative impact of the test's problems could be minimized. Of course, individual institutions and administrators might attempt to use the test's results in a way that is not in keeping with the CAE's own documentation. But faculty would have a powerful argument against doing so, thanks to the repeated insistence within the CLA+'s research and documentation that the test is not to be used this in this manner.

The test also has considerable weaknesses, as discussed at length in Chapter 3. But those weaknesses are also generally present in similar tests that may be used in its place. For example, the motivation issue is an identified problem in any of the three potential critical thinking tests to be utilized in Purdue's assessment initiative. Issues of attrition are common to all longitudinal tests; ceiling effects, a potential problem with all educational testing. In other words, the CLA+ has decided advantages compared to other

tests, while its significant drawbacks are generally ones shared by other tests as well. Again, the question is one of institutional and political realism: if some sort of test is inevitable, we as members of a college community should identify and advocate for least-bad options.

Accountability Cuts Both Ways

One of the enduring motifs of discussions of educational assessment is the need for accountability. That is, educators and educational institutions are argued to have a responsibility to demonstrate the effectiveness of their efforts to various stakeholders, such as students, parents, and the taxpayers who partially fund schools and universities. This call is coming from the heights of our educational policy apparatus. “As a nation, we have to make college more accessible and affordable and ensure that all students graduate with a quality education of real value," said Education Secretary Arne Duncan recently. "Our students deserve to know, before they enroll, that the schools they've chosen will deliver this value” (Bidwell). Tests such as the CLA+ are intended to make learning outcomes publicly available, and ideally to make the results accessible and interpretable for as broad an audience as possible. In other words, a key element of the accountability that assessment efforts are meant to promote is *transparency*, open access to information that can ultimately improve education for all involved.

Yet the process of researching and writing this dissertation demonstrates the degree to which transparency of assessment systems themselves can’t be assumed. This research originally was conceived of as an investigation of actual CLA+ student responses, comparing these texts to student scores to find patterns in how the test operates. However, I was unable to obtain a data set with which to effect this analysis, as

the CAE as a matter of policy does not provide student responses on the CLA+ to anyone—even the institutions that commission the test. Institutions have no ability to see their actual student output, despite paying $35 a student. In fact, the only CLA+ texts that are made available are a small handful of model responses that demonstrate how the test works. While representatives of the CAE recommended research literature for this project, the overall tenor of their responses was guarded. Haswell reflects on this tendency in his review of *Academically Adrift*, writing "Readers of AA are not allowed to know the actual prompts used or genres required—CLA test security. Nor can readers know what, if any, writing skills entered into that unitary CLA score… how these ratings were cross-validated or translated into a final number is a holistic mystery known only to CLA and the raters, who by contract can never tell" (489). This reticence stands in contrast to the CAE's explicit aims of increased transparency in educational testing. This is part of a broader tendency in educational testing for developers to maintain secrecy about their tests and how they work.

In his book *Measuring College Learning Responsibly,* Shavelson invokes accountability as the core justification for developing and implementing the CLA. "The notion of accountability makes clear that it is reasonable to expect public and private higher-education institutions to be held accountable to the public," writes Shavelson. "'Trust me' is an inadequate response to the demand for accountability" (123). This is a reasonable point of view, and as I have said from the outset of this project, I do not uniformly oppose higher education assessment efforts generally or the CLA+ specifically. But Shavelson's attitude demands a natural response: where, exactly, does accountability come from for the CAE and the test it develops? Who is checking up on their work in the

way that they would check up on colleges and their faculties? The CAE has generated a great deal of internally funded and directed research on the CLA and CLA+. To their credit, this internal research includes repeatedly soliciting and publishing critical research on the test. But surely they must recognize that any internally-generated research is subject to critique, given the clear conflict of interest inherent to such research. That natural conflict of interest does not mean that this research lacks value, but it does necessitate that such research be balanced with appropriately skeptical investigations developed by outside scholars. At present, the extant outside literature on the CLA+ is not remotely sufficient given the very high-stakes nature of its purpose.

This lack of transparency carried over into my efforts to obtain information from Purdue University stakeholders. While I was ultimately able to compile the necessary information to complete this document, and many members of Purdue's community were helpful in that regard, many involved in the assessment effort declined to participate in my research. This includes Dale Whitaker, the co-chair of the Student Growth Task Force; Mitch Daniels, the president of the university (despite initial indications from his office that he would be made available for an interview); Patricia Hart, the chairperson of the University Senate; Kirk Alter, a member of the SGTF Oversight Committee and a key figure in faculty resistance to the CLA+ initiative; Frank Dooley, Vice Provost for Undergraduate Academic Affairs; and others. This refusal made information collection more difficult, and ultimately left key perspectives outside of the scope of this research. Though none of these figures were under any obligation to participate, given that Purdue was a site of considerable disagreement over these issues and attracted national attention as such, the reticence is puzzling. A dissertation is unlikely to attract much attention or

readership beyond the committee to which it is submitted, but there is value in simply stating for the record one's individual take on an issue of such controversy. Particularly frustrating is that several of the potential interviewees that I contacted spoke to the media about this topic but not to a doctoral student within their own institution. Ultimately, I was able to assemble the information necessary to create a comprehensive local history of the CLA+ initiative at Purdue, but the number of interviews I conducted was a small fraction of my original intention.

At both the national level and the local level, the lack of transparency is disturbing, given the stakes involved and the continued invocation of accountability rhetoric. The lack of access to key information from test developers and the administrators who implement their products lends credence to long-standing complaints that the higher education reform movement is in fact a politicized privatization scheme, intended to strip faculty of control of their institutions and allow for greater influence of for-profit entities.

Writing Studies Must Adapt to Thrive

Since I have become part of the community of writing studies as a graduate student, I have considered myself part of a movement within the field calling for two aspects of our research that have fallen out of fashion: one, the use of empiricism, broadly defined, to investigate writing and writers; and two, research that takes as its primary interest student prose as product and process, in a limited sense. That is, I join with scholars like Haswell, Davida Charney, Susan Peck MacDonald, Keith Rhodes, Monica McFawn Robinson, and others in saying that the field of writing studies risks losing its disciplinary legitimacy, as well as its status as a research field entirely, if it does

not produce sufficient empirical work and pedagogical work to satisfy the universities in which it is embedded. I do not believe and have never claimed that all scholars within the field should undertake this work, or that this type of knowledge is more important or valid than others. In fact, I believe that doing more empirical and pedagogical work could help the field defend its institutional autonomy and thus the more theoretical and political work that it produces in great volume in its most prestigious journals.

These two facets of scholarship—prose pedagogy and empiricism— are precisely those that contribute most directly and usefully to the world of high stakes assessment. This dissertation cites very little research on standardized testing that comes specifically from the field of writing studies, simply because very little research of direct relevance has been produced within the field. It's therefore of little surprise to find that writing researchers have played small role, if any, in the development of tests like the CLA+. While I recognize that the quantitative and scientistic world of educational testing is unlikely to embrace writing studies and its humanist beliefs even under the best of times, I do believe that it is possible for members of our community to engage productively in the development of assessment systems, whether national or local, if we approach such potential collaborations with our best rhetoric. If any field should be able to speak the language of power to power in order to secure our influence and our future, it should be a field that has embraced rhetoric like ours has. Mutual suspicion serves no one, and keeping our heads in the sand will only hasten the assault on writing and the humanities as a field of research inquiry.

In the weeks prior to the completion of this dissertation, I inspired a long debate on the Writing Program Administrators listserv, a community of scholars that includes

many of the field's biggest lights. I expressed my concerns over the field's paucity of empirical research, and noted my fear that moving further and further away from prose pedagogy in the traditional sense leaves us vulnerable to hostile administrators in a time of declining funding for the humanities. A lively debate ensued, one which came to involve some of the biggest names in the field, such as Kathleen Blake Yancey, Victor Vitanza, Richard Haswell, and Chuck Bazerman. Bazerman expressed a sentiment very close to my own:

> I have been, as you may know, a committed advocate for research of many kinds in our field as well as a purveyor of theories that bear on writing, though not necessarily those that are usually identified as composition or rhetorical theory. Yet that research and theory has more vitality and creativity, I believe, when it keeps in mind the object of our profession. At the very least, public support for our profession depends on the perception that we are delivering on improving writing, and that we are producing knowledge that will lead to the improvement of writing. If our knowledge and practices are not perceived as improving writing (however that defined by relevant audiences), there are other groups ready to step in to claim the space (whether other academic disciplines or publishers or technology corporations). If we have no new knowledge but only commodified practice, we remain lowly paid, poorly trained deliverers of services. Only if we have relevant research and expanding knowledge that improves our professional practice can we thrive as a profession. ("Re: writing pedagogy/"the essay")

It is precisely this concern that motivates the research contained in this dissertation—the fear that, should we in writing studies fail to meaningfully provide to our institutions the teaching and research they define as our purpose, they will hand control of writing instruction over to others who are more willing to give them what they want. There is still the possibility that the higher education assessment movement may be a passing fad. Administrations change, controversies fade out, the public and politicians turn their attention elsewhere. Another test might become the dominant assessment of college learning, and the CLA+ might become irrelevant. But the deeper concern, of a field called writing studies that devotes so little of its research energy to writing in the traditional sense, will endure. These are dangerous, difficult times, not just for writing studies, not just for the humanities, but for the entirety of the academy. I do believe that we can react to these challenges and survive, even thrive, but we can only do so if we are willing to look at the world outside our windows, and change.

BIBLIOGRAPHY

BIBLIOGRAPHY

"About Dr. Cordova." *Purdue.edu*. Purdue University, nd. Web. 1 September 2014.

"About the GRE Subject Tests." *ETS.org*. Educational Testing Service, nd. Web. 5 July 2014.

"Accreditation in the United States." *Ed.gov.* U.S. Department of Education, nd. Web. 14 August 2014.

"Accreditation of Purdue University's West Lafayette Campus by the North Central Association of Colleges and Schools." *Purdue.edu*. North Central Association of Colleges and Schools, 1999. Web. 14 August 2014.

Alexander, Lamar. "'Time for Results': An Overview." *Phi Delta Kappan* (1986): 202-204. Print.

Anderson, Nick. "College Presidents on Obama's Rating Plan." *Washington Post* (2013): 08-31. Print.

Ansary, Tamim. "Education at risk: Fallout from a flawed report." *Edutopia.com*. Edutopia, 9 March 2007. Web. 15 May 2014.

Arum, Richard, and Josipa Roksa. *Academically adrift: Limited learning on college campuses*. University of Chicago Press, 2011. Print.

"Average Rates of Growth of Published Charges by Decade." *Trends in Higher Education.* The College Board, January 2014. Web. 19 January 2014.

Baker, Eva L., et al. "Problems with the Use of Student Test Scores to Evaluate Teachers. EPI Briefing Paper# 278." *EPI.org.* Economic Policy Institute, 2010. Web. 1 August 2014.

Bangert, Dave. "Daniels, Purdue faculty in test of wills." *Lafayette Journal & Courier.* 27 January 2015. Print.

Banta, Trudy and Steven Pyke. "Making the Case Against—One More Time." *The Seven Red Herrings About Standardized Assessments in Higher Education*. Roger Benjamin, ed. NILOA Occasional Paper. September 2012. Web. 19 January 2014.

Bazerman, Charles. "Re: writing pedagogy/'the essay.'" Message to the Writing Programs Administration Listserv. 28 March 2015. Email.

Benjamin, Roger, and Marc Clum. "A New Field of Dreams: The Collegiate Learning Assessment Project." *Peer Review* 5.4 (2003): 26-29. Print.

Benjamin, Roger, and Richard H. Hersh. "Measuring the difference college makes: The RAND/CAE value added assessment initiative." *Peer Review* 4.2/3 (2002): 7-11. Print.

Berlin, James A. *Writing instruction in nineteenth-century American colleges*. Carbondale, Illinois: SIU Press, 1984. Print.

Berlin, James A., and Michael J. Vivion. *Cultural Studies in the English Classroom*. Portsmouth, NH: Boynton/Cook, 1992. Print.

"Beyond the rhetoric: Improving college readiness through coherent state policy." *Southern Regional Education Board.* The National Center for Public Policy and Higher Education, June 2010. Web. 20 February 2015.

Bhattacharya, Tithi and Bill V. Mullen. "What's the Matter with Indiana?" *Jacobin Magazine,* 8 August 2013. Web. 1 September 2014.

Bidwell, Allie. “Obama Administration Seeks Input on College Ratings Draft.” *US News & World Report.* US News and World Report, 19 December 2014. Web. 25 December 2014.

Board of Trustees of Purdue University. “Trustees Elect Mitchell Daniels President of Purdue.” Message to the campus. 22 June 2013. E-mail.

Bound, John, and Sarah Turner. "Going to war and going to college: Did World War II and the GI Bill increase educational attainment for returning veterans?." *Journal of Labor Economics* 20.4 (2002): 784-815. Print.

Braun, Henry I. "Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models. Policy Information Perspective." *ETS.org.* Educational Testing Service, 2005. Web. 25 July 2014.

Brereton, John C. *The origins of composition studies in the American college, 1875-1925: A documentary history*. Pittsburgh: University of Pittsburgh Press, 1995. Print.

Card, David, and Thomas Lemieux. "Going to college to avoid the draft: The unintended legacy of the Vietnam War." *American Economic Review* (2001): 97-102. Print.

Celis, William. "Computer admissions test found to be ripe for abuse." *New York Times* (16 December 1994). Print.

Charney, Davida. "Empiricism is not a four-letter word." *College Composition and Communication* 47.4 (1996): 567-593. Print.

“CLA+ Rubric.” *CAE.org*. Council for Aid to Education, nd. Web. 22 January 2014.

“CLA+ Sample Tasks.” *CAE.org*. Council for Aid to Education, nd. Web. 22 January 2014.

“CLA+ Task Format.” *CAE.org*. Council for Aid to Education, nd. Web. 22 January 2014.

“Comparing CLA to CLA+.” *CAE.org*. Council for Aid to Education, nd. Web. 22 January 2014.

Core Curriculum Committee. “Motion to Approve the Core Curriculum.” Purdue University, 2012. Print.

Costa, Arthur L., and Richard A. Loveall. "The Legacy of Hilda Taba." *Journal of curriculum and supervision* 18.1 (2002): 56-62. Print.

Council for Aid to Education. “Fall 2014 CLA+ Mastery Results: Purdue University.” January 2015. Electronic Media.

—.“Statement on the Role of Assessment in Higher Education.” *CAE.org.* Council for Aid to Education, nd. Web. 19 January 2014.

Daiker, Donald A., Jeff Sommers, and Gail Stygall. "The pedagogical implications of a college placement portfolio." *Assessment of writing: Politics, policies, practices.* New York: Modern Language Association of America, 1996: 257-270. Print.

Daniels, Mitchell. “A Message from President Daniels on the Gallup-Purdue Index.” *Purdue.edu.* Purdue University Office of the President. 17 December 2013. Web. 15 January 2014.

— “A Message from President Daniels About Tuition.” *Purdue.edu.* Purdue University Office of the President. 18 March 2013. Web. 15 January 2014.

— “Student Growth Task Force Memo.” 6 March 2013. Print.

Daniels, Mitchell and Tim Sands. “A Message from President Daniels and Provost Sands about Foundations of Excellence.” *Purdue.edu*. Purdue Universty, 25 March 2014. Web. 20 March 2015.

“Does College Matter?” *CAE.org.* The Council for Aid to Education, January 2013. Web. 22 January 2014.

Educational Testing Service. “General Education Assessments: ETS Pilot With Purdue University.” 2 September 2014. Electronic Media.

"Education Pays: More education leads to higher earnings, lower unemployment." *bls.gov* National Bureau of Labor Statistics, 2009. Web. 21 May 2014.

"Estimating the Effect a Ban on Racial Preferences Would Have on African-American Admissions to the Nation's Leading Graduate Schools." *The Journal of Blacks in Higher Education* 19 (Spring, 1998): 80-82. Print.

"E-Update: The GRE Revised General Test." *ETS.org*. Educational Testing Service, May 2010. Web. 5 July 2014.

Ewing, John. "Mathematical intimidation: Driven by the data." *Notices of the AMS* 58.5 (2011): 667-673. Print.

"Fact sheet on the president's plan to make college more affordable: A better bargain for the middle class." *WhiteHouse.gov*. United States Government, 22 August 2013. Web. 1 July 2014.

"Fast Facts: Enrollment." *NCES.gov*. U.S. Department of Education Institute of Education Sciences National Center for Education Statistics, 2013. Web. 31 March 2015.

Flaherty, Colleen. "Test Anxiety." *Inside Higher Ed.* Inside Higher Ed, 28 January 2015. Web. 29 January 2015.

Flynn, Elizabeth A. "Feminism and scientism." *College Composition and Communication* 46.3 (1995): 353-368. Print.

"Forty-nine states and territories join Common Core Standards Initiative." *NGA.org.* National Governors Association, 1 June 2009. Web. 5 February 2015.

"Foundations of Excellence." *Purdue.edu.* Purdue University, nd. Web. 14 August 2014.

"Frequently Asked Technical Questions." *CAE.org*. The Council for Aid to Education. 2012. Web. 30 January 2014.

Fulcher, Glenn. *Practical language testing*. Abingdon: Routledge, 2013. Print.

Fulkerson, Richard. "Composition at the turn of the twenty-first century." *College Composition and Communication* 56.4 (2005): 654-687. Print.

Fulmer, Sara M., and Jan C. Frijters. "A review of self-report and alternative approaches in the measurement of student motivation." *Educational Psychology Review* 21.3 (2009): 219-246. Print.

Gallagher, Chris W. "Review Essay: All Writing Assessment is Local." *College Composition and Communication* 65.3 (February 2014): 486-505. Print.

Gardner, David P. *A Nation at Risk*. Washington, DC: The National Commission on Excellence in Education, US Department of Education, 1983. Print.

George, Diana, and John Trimbur. "Cultural studies and composition." *A guide to composition pedagogies* (2001): 71-91. Print.

Guthrie, James W., and Matthew G. Springer. "A Nation at Risk Revisited: Did" Wrong" Reasoning Result in" Right" Results? At What Cost?." *Peabody Journal of Education* 79.1 (2004): 7-35. Print.

Hannah, Susan B. "The higher education act of 1992: Skills, constraints, and the politics of higher education." *The Journal of Higher Education* (1996): 498-527. Print.

Haswell, Richard. "Methodologically Adrift." *College Composition and Communication* 63.3 (February 2012): 487-491. Print.

—. "NCTE/CCCC's recent war on scholarship." *Written Communication* 22.2 (2005): 198-223. Print.

Haynie, Devon. "What Online Students Need to Know About Automated Grading." *US News and World Report*. US New and World Report, 13 June 2014. Web. 18 June 2014.

Herndl, Carl G. “Teaching Discourse and Reproducing Culture: A Critique of Research and Pedagogy in Professional and Non-Academic Writing.” *College Composition and Communication* 44.3 (Oct., 1993): 349-363. Print.

Hibel, Thomas. “What Does the History of Faculty Unions Teach Us About Their Future?” *HigherEdJobs.com*. HigherEdJobs, nd. Web. 1 March 2015.

“History.” *CAE.org*. The Council for Aid to Education, nd. Web. 15 May 2014.

Hosch, Braden J. "Time on test, student motivation, and performance on the Collegiate Learning Assessment: Implications for institutional accountability." *Journal of Assessment and Institutional Effectiveness* 2.1 (2012): 55-76. Print.

Hunter, John E., and Hunter, Ronda F. Validity and utility of alternative predictors of job performance. *Psychological Bulletin* 96.1 (Jul 198): 72-98. Print.

Huot, Brian. *(Re)Articulating Writing Assessment for Teaching and Learning*. Logan, Utah: Utah State Press, 2002. Print.

Hursh, David W. *High-stakes testing and the decline of teaching and learning: The real crisis in education*. Vol. 1. Rowman & Littlefield, 2008. Print.

“International Students and Scholars.” *Purdue.edu.* Purdue University, nd. Web. 14 August 2014.

Jaschik, Scott. “Tests With and Without Motivation. *Inside Higher Ed*. Inside Higher Ed, 2 January 2013. Web. 20 July 2014.

Jastrzebski, Stan. “Can A Test Show How Purdue Students Grow? And Is It Even Worth It?” *WBAA.* WBAA, 4 March 2015. Web. 4 March 2015.

Kaplan, Robert, and Dennis Saccuzzo. *Psychological testing: Principles, applications, and issues*. Cengage Learning, 2012. Print.

Kelter, Laura A. "Substantial job losses in 2008: Weakness broadens and deepens across industries." *Monthly Labor Review* 132 (2009): 20. Print.

Kiley, Kevin. “Where Universities Can Be Cut.” Inside Higher Ed. InsideHigherEd.com, 16 September 2011. Web. 28 November 2014.

Kingkade, Tyler. “Mitch Daniels Protested As Purdue University's Next President By Students, Faculty, Alumni.” *Huff Po College.* Huffington Post, 2 July 2014. Web. 20 August 2014.

Klein, Stephen P., et al. "An Approach To Measurng Cognitive Out Comes Across Higher Education Institutions." *Research in Higher Education* 46.3 (2005): 251-276. Print.

—. "The Collegiate Learning Assessment : Facts and Fantasies." *Evaluation Review* 31.5 (2007): 415-439. Print.

—. "Test Validity Study (TVS) Report." *CAE.org*. The Council for Aid to Education & Fund for the Improvement of Postsecondary Education. 29 September 2009. Web. February 2 2014.

Layton, Lyndsey. "“How Bill Gates pulled off the swift Common Core revolution." *Washington Post* 7 June 2014. Print.

Lederman, Doug. “18 Yesses, One Major No.” *Inside Higher Ed.* Inside Higher Ed, 11 August 2006. Web. 15 May 2014.

Liu, Ou Lydia, Brent Bridgeman, and Rachel M. Adler. "Measuring learning outcomes in higher education motivation matters." *Educational Researcher* 41.9 (2012): 352-362. Print.

McCaffrey, Daniel F., et al. *Evaluating Value-Added Models for Teacher Accountability*. Santa Monica: RAND Corporation, 2003. Print.

Meagher, Michael E. "" In an Atmosphere of National Peril": The Development of John F. Kennedy's World View." *Presidential Studies Quarterly* (1997): 467-479. Print.

Miller, Susan. *Textual carnivals: The politics of composition*. Carbondale, Illinois: SIU Press, 1993. Print.

“Mitch E. Daniels Biograph.” *Purdue.edu*. Purdue University, nd. Web. 1 September 2014.

“Morrill Act.” *Library of Congress.* Library of Congress, nd. Web. 11 August 2014.

Moss, Pamela A. "Testing the test of the test: A response to “multiple inquiry in the validation of writing tests”." *Assessing Writing* 5.1 (1998): 111-122. Print.

“National University Rankings.” *US News and World Report.* US News and World Report, 2014. Web. 11 August 2014.

North, Stephen M. *The making of knowledge in composition: Portrait of an emerging field*. Upper Montclair, NJ: Boynton/Cook Publishers, 1987. Print.

Nystrand, Martin, Allan S. Cohen, and Norca M. Dowling. "Addressing reliability problems in the portfolio assessment of college writing." *Educational Assessment* 1.1 (1993): 53-70. Print.

Oliff, Phil, Vincent Palacios, Ingrid Johnson, and Michael Leachman. “Recent Deep State Higher Education Cuts May Harm Students and the Economy for Years to Come.” Center on Budget and Policy Priorities, 2013. Web. 15 January 2014.

Olson, Keith W. "The GI Bill and higher education: Success and surprise." *American Quarterly* (1973): 596-610. Print.

“Overview of Accreditation.” *Ed.gov.* U.S. Department of Education, nd. Web. 14 August 2014.

Paul, Joseph. “Purdue pauses standardized testing roll out.” *Lafayette Journal & Courier.* 9 April 2015. Web. 14 April 2015.

Peterson, M., Vaughan, D., & Perorazio, T. *Student assessment in higher education: A comparative study of seven institutions*. Ann Arbor: University of Michigan, National Center for Postsecondary Improvement, 1999. Print.

"Purdue University Accreditation." *Purdue.edu.* Purdue University,

"Purdue University—Main Campis." *Carnegie Classification of Institutes of Higher Education.* Carnegie Foundation for the Advancement of Teaching, nd. Web. 11 August 2014.

"Reliability and Validity of the CLA+." *CAE.org*. The Council for Aid to Education, 2012. Web. 30 January 2014.

"Remarks by the President on College Affordability, Ann Arbor, Michigan." *WhiteHouse.gov*. Office of the Press Secretary, 27 January 2012. Web. 2 June 2014.

Rhodes, Keith, and Monica M. Robinson. "Sheep in Wolves' Clothing: How Composition's Social Construction Reinstates Expressivist Solipsism (And Even Current-Traditional Conservatism)." *The Journal of the Assembly for Expanded Perspectives on Learning* 19.1 (2013): 8-22. Print.

Scharton, Maurice. "The politics of validity." *Assessment of writing: Politics, policies, practices.* New York: The Modern Language Association of America (1996). Print.

Second Report: President's Committee on Education Beyond the High School. Washington: United States Government, 1956. Print.

Selingo, Jeff. "President Sees an Obamacare Solution to Higher Ed's Problems." *Chronicle of Higher Education.* Chronicle of Higher Education, 22 August 2013. Web. 1 July 2014.

Shavelson, Richard. *Measuring college learning responsibly: Accountability in a new era*. Stanford University Press, 2009. Print.

Shermis, Mark D. "The Collegiate Learning Assessment: A Critical Perspective." *Assessment Update* 20.2 (2008): 10-12. Print.

Shermis, Mark D., and Jill C. Burstein, eds. *Automated essay scoring: A cross-disciplinary perspective.* Mahwah, New Jersey: Lawrence Erlbaum Associates, 2003. Print.

Sloan, John W. "The management and decision-making style of President Eisenhower." *Presidential Studies Quarterly* (1990): 295-313. Print.

Spellings, Margaret. *A Test of Leadership: Charting the Future of US Higher Education*. United States Department of Education.: 2006. Print.

"Standing Committees." *Purdue.edu.* Purdue University, nd. Web. 20 March 2015.

Stark, Joan S., and Lisa R. Lattuca. *Shaping the college curriculum: Academic plans in action*. Boston: Allyn and Bacon, 1997. Print.

"State by State Data: Indiana." *The Project on Student Debt.* The Institute for College Access and Success, 2014. Web. 17 March 2015.

"Statement by the President Making Public a Report of the Commission on Higher Education." *The American Presidency Project.* University of California—Santa Barbara, nd. Web. 15 June 2014.

Steedle, Jeffrey T. "Incentives, motivation, and performance on a low-stakes test of college learning." *Annual Meeting of the American Educational Research Association*. Denver, CO: 2010. Print.

Sternberg, Robert J., and Wendy M. Williams. "Does the Graduate Record Examination predict meaningful success in the graduate training of psychology? A case study." *American Psychologist* 52.6 (1997): 630. Print.

Stratford, Michael. "Margaret Spellings on Obama Plan." *Inside Higher Ed.* Inside Higher Ed, 5 September 2013. Web. 30 Jun 2014.

"Student Debt and the Class of 2012." *Project On Student Debt.* Institute for College Access and Success, December 2013. Web. 19 January 2013.

Student Growth Task Force Committee. *Understanding and Fostering Student Growth: A Purdue University Commitment.* November 2013. Print.

Student Growth Task Force Oversight Committee. *Report of the Student Growth Task Force Oversight Committee.* December 2014. Print.

Taba, Hilda, and Willard B. Spalding. *Curriculum development: Theory and practice*. New York: Harcourt, Brace & World, 1962. Print.

Teddlie, Charles. *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. Thousand Oaks, CA: Sage Publications Inc, 2009. Print.

"Trends in College Pricing 2014." *Trends in Higher Education Series.* The College Board, 2014. Web. 15 March 2015.

Tyler, Ralph W. *Basic principles of curriculum and instruction*. Chicago: University of Chicago Press, 2013. Print.

"Use of CAE Data." *CAE.org*. The Council for Aid to Education. 2012. Web. 30 January 2014.

Weisberg, Lauren. "Study gives Purdue's core curriculum a failing grade." *Purdue Exponent.* Purdue Exponent, 21 May 2012. Web. 11 November 2014.

Weissmann, Jordan. "How Bad is the Job Market for the Class of 2014?" *Slate.com*. Slate, 8 April 2014. Web. 15 June 2014.

"What is the Value of the Graduate Record Examinations?" *ETS.org.* Educational Testing Service, October 2008. Web. 14 July 2014.

Whittaker, Dale. "Email from Dale Whittaker." Message to the University Senate. 1 April 2014. Email.

White, Edward M. *Teaching and Assessing Writing: Recent Advances in Understanding, Evaluating, and Improving Student Performance*. San Francisco: Jossey-Bass, 1994. Print.

—."Holistic scoring: Past triumphs, future challenges." *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*. Cresskill, NJ:

White, Edward Michael, William Lutz, and Sandra Kamusikiri, eds. *Assessment of writing: Politics, policies, practices*. New York: Modern Language Association of America, 1996. Print.

Winterbottom, John A. "The Graduate Record Examination Program." *ADE Bulletin* 20 (1969): 43-49. Print.

Wright, Barbara D. "Accreditation and the scholarship of assessment." *Building a scholarship of assessment*. San Francisco: Jossey-Bass, 2002. 240-258. Print.

Writing Assessment in the 21st Century: Essays in Honor of Edward M. White. Norbert Elliot and Les Perelman, eds. New York: Hampton, 2012. Print.

Yancey, Kathleen Blake. "From the Editor: Speaking Methodologically." *College Composition and Communication* 64.1 (September 2012): 5-14. Print.

—. "Looking back as we look forward: Historicizing writing assessment." *College Composition and Communication* 50.3 (1999): 483-503. Print.

Zemsky, Robert. "The unwitting damage done by the Spellings Commission." *Chronical of Higher Education*. Chronicle of Higher Education, 18 September 2011. Web. 20 June 2015.

Zook, George F. *Higher education for American democracy, a report*. Washington: United States Government, 1947. Print.

APPENDICES

Appendix A IRB Application

Revised 10/10

Re

f. # _____

APPLICATION TO USE HUMAN RESEARCH SUBJECTS

Purdue University
Institutional Review Board

1. Project Title: The CLA+ and the Two Cultures: Writing Assessment and Educational Testing

2. Full Review ☐ Expedited Review ☒

3. Anticipated Funding Source: No Funding Necessary

4. Principal Investigator *[See Policy on Eligibility to serve as a Principal Investigator for Research Involving Human Subjects]*:
   Richard Johnson-Sheehan, Professor English, Heavilon, (765) 494-3740, (765) 494-3780, rjohnso@purdue.edu

5. Co-investigators and key personnel *[See Education Policy for Conducting Human Subjects Research]*:
   Fredrik deBoer, PhD Candidate English, Heavilon, (860) 336-9931, no FAX, fdeboer@purdue.edu

6. Consultants *[See Education Policy for Conducting Human Subjects Research]*:
   None Department, Building, Phone, FAX, E-mail address

7. The principal investigator agrees to carry out the proposed project as stated in the application and to promptly report to the Institutional Review Board any proposed changes and/or unanticipated problems involving risks to subjects or others participating in the approved project in accordance with the HRPP Guideline 207 Researcher Responsibilities, Purdue Research Foundation-Purdue University Statement of Principles and the Confidentiality Statement. The principal investigator has received a copy of the Federal-Wide Assurance (FWA) and has access to copies of 45 CFR 46 and the Belmont Report. The principal investigator agrees to inform the Institutional Review Board and complete all necessary reports should the principal investigator terminate University association.

_____

_____

Principal Investigator Signature Date

8. The Department Head (or authorized agent) has read and approved the application. S/he affirms that the use of human subjects in this project is relevant to answer the research question being asked and has scientific or scholarly merit. Additionally s/he agrees to maintain research records in accordance with the IRB's research records retention requirement should the principal investigator terminate association with the University.

_____

_____

Department Head *(printed)* Department Name

_____

_____

Department Head Signature Date

- **APPLICATION TO USE HUMAN RESEARCH SUBJECTS**

9. This project will be conducted at the following location(s): (please indicate city & state)

☒ Purdue West Lafayette Campus
☐ Purdue Regional Campus (Specify): _____
☐ Other (Specify): _____

10. If this project will involve potentially vulnerable subject populations, please check all that apply.

☐ Minors under age 18
☐ Pregnant Women
☐ Fetus/fetal tissue
☐ Prisoners Or Incarcerated Individuals
☐ University Students (PSYC Dept. subject pool ___)
☐ Elderly Persons
☐ Economically/Educationally Disadvantaged Persons
☐ Mentally/Emotionally/Developmentally Disabled Persons
☐ Minority Groups and/or Non-English Speakers
☐ Intervention(s) that include medical or psychological treatment

11. Indicate the anticipated maximum number of subjects to be enrolled in this protocol as justified by the hypothesis and study procedures: _____12_____

12. This project involves the use of an **Investigational New Drug** (IND) or an **Approved Drug For An Unapproved Use**.

☐ YES ☒ NO

Drug name, IND number and company: _____

13. This project involves the use of an **Investigational Medical Device** or an **Approved Medical Device For An Unapproved Use**.

☐ YES ☒ NO

Device name, IDE number and company: _____

14. The project involves the use of **Radiation or Radioisotopes**:

☐ YES ☒ NO

15. Does this project call for: (check-mark all that apply to this study)

☒ Use of Voice, Video, Digital, or Image Recordings?
☐ Subject Compensation? Please indicate the maximum payment amount to subjects. $_____
Purdue's Human Subjects Payment Policy Participant Payment Disclosure Form
☐ VO2 Max Exercise?
☐ More Than Minimal Risk?
☐ Waiver of Informed Consent?

☐ Extra Costs To Subjects?
☐ The Use of Blood? Total Amount of Blood _____
Over Time Period (days) _____
☐ The Use of rDNA or Biohazardous materials?
☐ The Use of Human Tissue or Cell Lines?
☐ The Use of Other Fluids that Could Mask the Presence of Blood (Including Urine and Feces)?
☐ The Use of Protected Health Information (Obtained from Healthcare Practitioners or Institutions)?
☐ The Use of academic records?

16. Does investigator or key personnel have a potential financial or other conflict of interest in this study?
☐ YES ☒ NO

- **APPLICATION NARRATIVE**

A. PROPOSED RESEARCH RATIONALE

- The proposed research seeks to build a local history of the implementation of the Collegiate Learning Assessment+ (CLA+) here at Purdue University. Using a journalistic style, the co-investigator will interview prominent professional members of the campus community who are stakeholders in this assessment. These interviews will be utilized in building an oral history of how the push for a standardized assessment began at Purdue University, how the administrative and research teams involved in this effort were formed, how the CLA+ was chosen, and what impressions have been about the initial implementation of the CLA+ pilot program here at the university.

  Although some information has been made publicly available about the new assessment effort at Purdue University, there is much information that has yet to be revealed. In particular, how the CLA+ was chosen, what alternatives were considered, what the pros and cons of this particular assessment mechanism were considered to be, and what concerns or reservations were voiced by stakeholders involved in the process of selecting the CLA+. All of this information could be obtained by interviewing those stakeholders. Additionally, as Purdue is a diverse community with a variety of actors pursuing various ends and protecting certain educational and institutional values, there is value in reporting a variety of points of view in the implementation of this sort of major policy effort.

  The broader research rationale lies in the lack of current historical and sociological information about institution-level implementation of new assessment mechanisms within the American university. The current assessment push in our higher education system, most directly caused by the Spellings Commission Report, has been discussed in a number of

dissertations and academic articles. However, this research is almost exclusively focused on changes at the national or state level; these histories provide information about political and legal aspects of the assessment effort, but fail to demonstrate how this effort plays out in the local, institutional level. This project is an effort to build such a local history, and to tie that history into the larger national story of the recent assessment push, in a way that could provide guidance to other institutions and deepen our understanding of how institutions work to implement large-scale policy changes.

- **B. SPECIFIC PROCEDURES TO BE FOLLOWED**
  - Interviews will be conducted with research subjects to create a local history of the CLA+ and the assessment effort that led to its implementation at Purdue University. Questions will concern the pedagogical, institutional, administrative, legal, economic, and practical conditions, goals, and philosophies that contributed to the selection of this particular assessment instrument; controversies, disagreements, and concerns about the use of the assessment at Purdue University; impressions of the success of the early stages of implementation and piloting of the CLA+ mechanism; and predictions for the future of this assessment and how it will impact Purdue's community.
  - Interviews will be recorded with the informed consent of the research subjects. These interviews will be transcribed and analyzed by the co-investigator. Interview questions will be subject-specific. Questions will include both pre-scripted questions and questions that emerge during the interview, such as follow-up questions and questions for clarification. These interviews will be used to generate a timeline of the assessment initiative undertaken by the Daniels administration. They will help identify major players in this initiative and how they impacted the decision to use the CLA+. The interviews will clarify what process was undertaken to select the CLA+, what alternatives were explored, which aspects of the CLA+ were appealing, and what concerns were voiced during the process. They will also allow those involved in the CLA+ implementation to weigh in on controversies and criticisms related to the assessment push in general and to the CLA+ in particular.

C. SUBJECTS TO BE INCLUDED

Describe:

- This research concerns specific individuals who are professionally affiliated with Purdue University. They have been chosen because of their specific role in the selection of the CLA+, or because their given professional, institutional, academic, or administrative responsibilities at Purdue University make them stakeholders in the implementation of the assessment mechanism. No more than twelve (12) subjects will be interviewed. Potential interview subjects include, but are not limited to

- Mitch Daniels, President of Purdue University
- Dale Whittaker, Vice Provost for Undergraduate Academic Affairs
- Diane Beaudoin, Director of Assessment, Office of the Provost
- Sarah Bauer, Student Analytical Research
- Chantal Levesque-Bristol, Director of the Center for Instructional Excellence
- Jennifer Bay, Director of Introductory Composition at Purdue
- Patricia Hart, Chairperson of Purdue University Senate

D. RECRUITMENT OF SUBJECTS AND OBTAINING INFORMED CONSENT

- Each potential interviewee has been selected based on the nature of their position at Purdue University and their relationship to the implementation of the CLA+. Therefore recruitment will be a matter of seeking their informed consent individually.

E. PROCEDURES FOR PAYMENT OF SUBJECTS

- Subjects will not be paid for this research.

F. CONFIDENTIALITY

- Because of the nature of this research, and the specific selection of these individuals for their institutional positions and expertise, research subjects will not be anonymized.
- Research records, including audio files of interview and interview transcripts, will be stored only on local electronic storage such as a flash drive. This electronic local storage will be stored in a locked container when not being used by the investigators.
- There are no plans to destroy research records obtained in this research.

G. POTENTIAL RISKS TO SUBJECTS

- The risks associated with this research are minimal. There are potential professional or social risks for participants engaging in these interviews, given that the implementation of the CLA+ is an issue of potential controversy within the campus community. However, these risks are present in the commission of the daily duties and responsibilities of the research subjects, and are firmly in the control of the research subjects during their interviews.

H. BENEFITS TO BE GAINED BY THE INDIVIDUAL AND/OR SOCIETY

- The potential direct benefit of this research to research subjects is the ability to present their opinions and version of events in a local history of the implementation of the CLA+, and in so doing influence this project's understanding of how this development came to pass.
- The potential benefit for society is the creation of a local history of an assessment mechanism at a large public research university, one which will be connected to the public history of the larger assessment movement in the United States of the last decade. This history will provide researchers, faculty, administrators, and others with a deeper understanding of what the assessment movement means for the American university both nationally and locally, and demonstrate how an assessment system is implemented in a real university.

I. INVESTIGATOR'S EVALUATION OF THE RISK-BENEFIT RATIO

- As the potential benefits of this research is high, and the risks low and under the control of the research subjects, and the research subjects are all professional adults affiliated with Purdue University, the risk-benefit ratio seems to clearly fall in favor of performing this research.

J. WRITTEN INFORMED CONSENT FORM *(to be attached to the Application Narrative)*

- A written informed consent form is included in this submission.

K. WAIVER OF INFORMED CONSENT OR SIGNED CONSENT

If requesting either a waiver of consent or a waiver of signed consent, please address the following:

1. For a Waiver of Consent Request, address the following:
   a. Does the research pose greater than minimal risk to subjects (greater than everyday activities)?
   b. Will the waiver adversely affect subjects' rights and welfare? Please justify?
   c. Why would the research be impracticable without the waiver?
   d. How will pertinent information be reported to subjects, if appropriate, at a later date?

2. For a Waiver of Signed Consent, address the following:
   a. Does the research pose greater than minimal risk to subjects (greater than everyday activities)?
   b. Does a breech of confidentiality constitute the principal risk to subjects?
   c. Would the signed consent form be the only record linking the subject and the research?
   d. Does the research include any activities that would require signed consent in a non-research context?

e. Will you provide the subjects with a written statement about the research (an information sheet that contains all the elements of the consent form but without the signature lines)?

L. INTERNATIONAL RESEARCH

When conducting international research investigators must provide additional information to assist the IRB in making an appropriate risk/benefit analysis. Please consult the bullet points below when addressing this section of the application.

- Research projects must be approved by the local equivalent of an IRB before Purdue's IRB can grant approval to the protocol. If there is not equivalent board or group, investigators must rely on local or cultural experts or community leaders to provide approval and affirm the research procedures are appropriate for that culture. The Purdue IRB requires documentation to be submitted of this "local approval" before granting approval of the protocol. Additionally, please provide information about the IRB equivalent and provide contact information for the local entity. The body or individual providing the local approval should be identified in the application narrative as well as information as to that body's or individual's expertise.
- In the application narrative describe the experience and/or other qualifications the investigators have related to conducting the research with the local community/culture. Describe if the investigators have the knowledge or expertise of the local or state or national laws that may impact the research. The investigators must understand community/cultural attitudes to appreciate the local laws, regulations or norms to ensure the research is conducted in accordance with U.S. regulations as well as local requirements.
- For more information on specific requirements of different countries and territories, investigators can consult the Office for Human Research Protections International Compilation of Human Research Protections (http://www.hhs.gov/ohrp/international/). This is only one resource and it may not be an appropriate resource for your individual project.
- In the application narrative describe how the investigators will have culturally appropriate access to the community. If the investigators were invited into the community to conduct the research, please submit documentation of the collaboration.
- In the application narrative explain the investigators' ability to speak, read or write the language of potential participants. Describe the primary language spoken in the community. Explain provisions for culturally appropriate recruitment and consent accommodations translated materials or translators.
- Attention should be given to local customs as well as local cultural and religious norms when writing consent documents or proposing alternative consent procedures. This information should be provided in the application narrative, and as appropriate, provide justification if requesting the IRB to waive some or all requirements of written consent.
- In the application narrative describe how investigators will communicate with the IRB while you are conducting the research in the event the project requires changes

or there are reportable events. Also, if the researcher is a student, describe how the student will communicate with the principal investigator during the conduct of the research and how the principal investigator will oversee the research.

- If this research is federally funded by the United States, additional documentation and inter-institutional agreements may be required. Contact the IRB Administrator for assistance.
- Submit copies of consent documents and any other materials that will be provided to subjects (e.g., study instruments, advertisements, etc.) in both English and translated to any other applicable languages.

M. SUPPORTING DOCUMENTS ***(to be attached to the Application Narrative)***

- Recruitment advertisements, flyers and letters.
- Survey instruments, questionnaires, tests, debriefing information, etc.
- If the research is a collaboration with another institution, the institution's IRB or ethical board approval for the research.
- If the research accesses the PSYC 120 Subject pool include the description to be posted on the web-based recruitment program (formerly *Experimetrix*).
- Local review approval or affirmation of appropriateness for international research.
- If the research will be conducted in schools, businesses or organizations, include a letter from an appropriate administrator or official permitting the conduct of the research.

Appendix B Informed Consent Form

For IRB Use Only

RESEARCH PARTICIPANT CONSENT FORM

The CLA+ and the Two Cultures: Writing Assessment and Educational Testing

Richard Johnson-Sheehan
Department of English
Purdue University

- **What is the purpose of this study?**

- The purpose of this study is to build a local history of the implementation of the Collegiate Learning Assessment+ (CLA+) at Purdue University. Its goal is to investigate how the initiative to assess student learning of Purdue University undergraduates began, why the CLA+ was chosen as the tool for that assessment, what controversies or problems were involved in this implementation, and how the initial stages of this implementation have gone.
- This interview is part of research contributing to Fredrik deBoer's doctoral dissertation in partial fulfillment of the degree requirements for a PhD in English with a focus on rhetoric and composition.
- You are being asked to participate because your professional, institutional, or administrative position within Purdue University makes your experience and history with the implementation of the CLA+ at Purdue relevant to this study.
- The expected number of participants in these interviews is about 12.

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

- If you choose to be in this study, you will be interviewed orally for a period not exceeding an hour and a half and unlikely to exceed an hour. The interview will involve questions regarding your involvement with or reaction to the CLA+ at Purdue and questions about your general attitude towards assessment in postsecondary education. This interviews will be audio recorded and transcribed by the researcher. It is possible that follow-up questions may be asked via email if necessary.

How long will I be in the study?

- Your participation will last only as long as is necessary to conduct the oral interview, likely not exceeding an hour, and for how long it may take to respond to follow-up questions via email. This research is expected to be fully completed by May 2015.

- **What are the possible risks or discomforts?**

- The potential risks of this research are minimal. They may include social or professional discomfort caused by your responses to interview questions.
- Because you will control what you say during our interview, you will be able to determine your own level of these social or professional risks.
- The risks of this study are no greater than you would encounter in daily life research.

- **Are there any potential benefits?**

- There are no anticipated direct benefits to you as a research participant beyond the ability to affect the local history this research will build.
- There may be a benefit to the larger Purdue community in the development of a local history of this assessment measure, as well as a benefit to researchers and others in connecting such a local history to the broader understanding of assessment in higher education.

- **Will information about me and my participation be kept confidential?**

- The project's research records may be reviewed by departments at Purdue University responsible for regulatory and research oversight.

- Because of the nature of this research and the importance of interviewing direct stakeholders in the implementation of the CLA+ on campus, your participation in this study will not be made confidential and your identity will not be anonymized.
- The primary investigator Richard Johnson-Sheehan, the co-researcher Fredrik deBoer, and dissertation committee members April Ginther and Nathan Johnson will have access to this research. Additionally, a fourth member of the dissertation committee will be appointed and given access to this research.
- Both the audio files of these interviews and their transcriptions will be kept electronically in the possession of Fredrik deBoer, on a local storage device such as a flash drive. This flash drive will be kept in a locked cabinet when not being used.
- There are no plans to destroy the records of this research.
- The results of this study will be disseminated in the form of Fredrik deBoer's doctoral dissertation, which may later be published in whole or in part, and which may become available to outside readers through online research databases such as ProQuest Dissertation Search.
- The researchers involved in this study cannot guarantee that your responses to these questions will remain confidential.

- **What are my rights if I take part in this study?**

Your participation in this study is voluntary. You may choose not to participate or, if you agree to participate, you can withdraw your participation at any time without penalty or loss of benefits to which you are otherwise entitled.

- You may withdraw your participation from this study at any time by asking to end the interview. You may choose not to answer follow-up or preliminary questions via email.
- Your responses, once recorded and transcribed, will be eligible for inclusion in this research. You may access the recorded audio or written transcription of this interview on request.
- Your participation or refusal to participate in this research will have no impact on your employment or standing at Purdue University.

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 Dr. Richard Johnson-Sheehan at (765) 494-3740 or Fredrik deBoer at (860) 336-9931. Fredrik deBoer should be your first point of contact.

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 Interview Transcripts

Diane Beaudoin Interview Transcript

9/17/2014

Fredrik deBoer: OK. Can you tell me a little bit about your position here at the University vis a vis assessment.

Diane Beaudoin: Currently I am Director of Assessment, right now. For the last six months, that has sat within the new Office of Institutional Research, Assessment, and Effectiveness. So OIRAE. Prior to that, my position had been in the Provost's office. So doing assessment and accreditation for the university under the direction of the Provost's office. Six months ago, that got moved. So, primarily, when I was in the Provost's office it was assessment of student learning, assessment of academic programs, things like that, as well as helping with student success metrics-- four year graduation rates, retention, things like that. I served as the accreditation liaison officer for the university. So working with our university accreditation as well as helping individual colleges and programs with their specializing accreditation. With the new OIRAE office that formed six months ago, I keep doing everything I used to do, but now I've got twelve people that report to me who have been doing assessment across campus in areas like diversity and inclusion, student affairs, student success, housing and food services, things like that. Trying to bring assessment together-- more coordination, so we're not all reinventing the same wheel, surveying students 10,000 times with the same questions, stuff like that.

Fd: one of the themes that's been running through my interviews is the perception by people within the institution-- Brent Drak said that Purdue is like 12 private colleges with a loose affiliation as opposed to one university.

DB: Right.

Fd: So do you see this new OIRAE as an effort to consolidate assessment efforts?

DB: Right.

Fd: Tell me a little bit about the selection and implementation of the CLA+. Were you directly involved in choosing the test?

DB: So, I don't know how much Brent shared, but this fall we piloted three critical thinking exams with the incoming freshmen during BGR, of which CLA+ was one. So we looked at CLA+, we did the CAAP, and CCTSC-- that's the California Critical Thinking... something test! I forget what the other S is in there! (laughter) We went with the three... I think there was an initial sense that we were just going to do CLA+ and go with that. When that was presented to the University Senate, there was some nervousness that this hadn't been vetted fully by faculty across the campus, that only this small task force of 8 got a vote and so a whole senate committee was formed then to get buy-in and

feedback on it, and so their recommendation was that we pilot several things before deciding which was going to be our end-all. We'd want to see what results looked like, what kind of feedback we'd get from the instruments, and so forth.

Fd: So as I understand it, part of the purpose of the piloting is to cross-validate each other, to confirm that they're finding similar findings.

DB: That's right.

Fd: Can you tell me a little bit about the students who were able to be pulled from Boiler Gold Rush?

DB: Um, so what we did for each of the three, we took um stratified random samples and each college was represented, demographics, things like that, so for each of them we did a poll of 800 students each, we invited them to participate, offered them each a $5 gift card, and they were all told that they would be entered into a raffle, to win either a $1000, $500, or $250 Amazon gift card. That email went out inviting people to take it. From each, we ended up with 100, from each group.

FD: So 300 total.

DB: Right.

FD: My understanding is that, in terms of Dale Whittaker's role, the plan was to implement the CLA+ eventually as not exactly a census-level test, but as a wide-scale instrument. I understand that you're now trying to undertake piloting efforts on a smaller scale. Is the plan still to make this a widespread test, once everything is in place?

DB: I think... based on these results, and now we've just been invited-- and the President's office has accepted on our behalf-- ETS is coming out with a new critical thinking test that they're going to pilot in February-- so we will do a pilot in February for that exam. Once all four of these exams, we have the results and whatever, then the faculty will choose which one. So it may not be CLA+. And then it will be... I don't think we'll ever get to census level. But I think we will do significant over-sample of our student population. So probably, you know, we look at our incoming classes of 7500 students, I would say we would probably like to get 2000. Then we could represent all of the colleges and be able to break out by gender and ethnicity for each college. But not ever get down to a departmental level where we could say, "Your kids aren't doing as well." But at least try to do gender, ethnic groups, college level differences... We're looking, probably, at 2000 to make any kind of statistically responsible...

FD: I was talking to Brent Drake.... One thing about the CLA is that they have this criterion sampling philosophy. And one of the things that they repeatedly say in their

materials is that you can't disaggregate their data in order to see what major or department is doing the best job. I very much believe that they believe that. I also believe that it's procedurally convenient for them, because when you're trying to implement this test at universities with powerful faculty senates, it helps allay their fears.

DB: That's correct. And that's why they say, "well you only need 200 to represent your students." But if you look at an institution the size of Purdue, 200 just isn't a representative sample.

FD: So considering that there's still choices to be made about what test is going to be implemented here, what do you think any minimally effective standardized assessment of college student learning has to accomplish? What do you think are the basic kinds of information that it has to provide for you to be useful?

DB: With a lot of standardized tests, you get a number. "Your students scored an 84.6 average." Well what does that mean? OK, great, that's how we compared to the national average, we were an 84 before and our seniors scored 91. To me... big deal, who cares? If the results don't have enough information to make actionable changes in curriculum, I don't see the purpose whatsoever. So unless there's some sort of subscores, subscales, that can really indicate to faculty in their critical thinking areas-- your students are strong in areas X, Y, Z-- it looks like criterion W is where your students need help -- here are suggestions, and what that subscore means... suggested ideas for skills you would want to improve on... The lumped, aggregated, "I got a 540 on my SAT," what does that mean? What was I good at, what was I not good at? Because you tested a million things. Whatever we go with, I would hope that the reported results gives some indication of what those subscales actually mean.

FD: Part of the particular and unique difficulty of assessing college students, and comparing students from across different institutions, is that colleges work very hard to make sure that their student bodies aren't the same. When we're testing, we want to test similar populations, but elite colleges put in a lot of work to ensure that their student populations are in fact different from others. We might be able to compare Purdue to Indiana University, but to compare our students with Harvard or with Ivy Tech, there's just differences in the incoming population. The CLA is usually looked at with a "value added" philosophy. Do you think that a value added approach, testing freshman and seniors, is a feasible solution here at Purdue? And if not, what are some ways we could control for differences in incoming student population?

DB: To be honest, I don't believe in the value added approach at all. I've watched students take standardized tests, three years ago as part of our Voluntary System of Accountability. I had to get 200 freshmen and 200 seniors to take the Proficiency Profile. I invited students, and had to do it in a proctored computer lab. I would proctor and just

watch. I could sit and look in that room and tell you which ones were freshmen and which ones were seniors. Freshmen took the exam, they typically took between 45 and 60 minutes for a 60 minute test. You could see them scribbling notes on scratch paper. Seniors were typically done in 15 minutes. Click, click, click, done-- "can I have my $5?" Done. You're not going to see the value added because there's no incentive for the students to score well, take it seriously, so whatever scores you get.... I don't think they show value added. In terms of trying to compare us to other institutions, I think you see the same things. I think....

FD: Because I know that's something that President Daniels has specific said he wants, is to ascertain greater value for a Purdue education. It strikes me that from a standpoint of pure test theory, that's awfully tricky to do.

DB: And even like you said, it's hard to compare across institutions. You don't know which group of students, you don't even know their personalities-- the ones who take this kind of thing seriously. What incentives did IU provide compared to what we provided? It might involve giving out a course grade, compared to us just giving out a Starbucks gift card. You have no idea what their recruiting incentives, their testing protocols even were.

FD: Talking about this issue of student motivation... in the research literature, there's a lot of skeptical arguments that are, frankly, looking for problems. But this motivation issue strikes me as the major one. And in particular, how to create incentives for students when the test doesn't have real-world applicability like an SAT score does. So are their any proposals for how to generate student motivation long term? Does this worry you?

DB: Yeah, I think it's a huge problem. CLA+'s solution to this is that they say they give individual students some kind of certificate, badge, whatever that they can show to employers that says, "I got a blah blah blah on Critical Thinking." Employers don't look for stuff like that. If you go talk to companies, they don't care. So I really don't know what the final motivation of any kind of critical thinking test would be to any student. And I don't think that's a problem we'll ever solve... I mean, we can get indications of students who take, like, a GRE to get to grad school. But there you're looking at a defined, small slice-- you've already taken an elite that gets into Purdue, then you're taking another elite, the students who want to go to grad school. They're taking GREs, MCATs, whatever.... Anything else we want to give to students, we can't say "you have to score X to graduate from Purdue."

FD: The CAE wants the CLA+ to become, if you will, the SAT of college students. But your own piloting demonstrates the difficulty of that. It's not like ETS, ACT, and other test companies are going to just say, "OK, great, you guys take it." There's competition.

And until there's kind of a critical mass of students taking this test, there's no reason for employers to take this test seriously -- and thus no reason for students to.

DB: Right.

FD: My particular fear-- and particularly because of this Obama proposal to tie federal funding to college rankings, and using tests like the CLA a part of those rankings-- you can see this sour spot where it's low stakes for students, and super high stakes for institutions. And that is very worrisome for me.

DB: If you really want, to go back to your value added question, it's not going to be a critical thinking test. You have to go to employers and ask them-- how do our engineering graduates compared to those from the University of Wisconsin? What are our strengths, and what strengths do you see from our graduates? In the aggregate, why do you hire Purdue grads over other institutions? Because companies have their lists of schools, their go-to places? We need to be talking to those companies and finding out why they keep coming back to hire our students over other students.

FD: Kind of an outcomes based assessment.

DB: Right.

FD: So ETS is coming this spring to implement their test this spring, and I'm sure they want you to choose theirs, right? It's very interesting to me to think about how US institutions have to critically evaluate these tests because in a very real sense their selling the tests to you. You said that the faculty senate will ultimately make a decision, is that right?

DB: A task force, yeah.

FD: In three years from now, what would you like to see in terms of broad testing at the university? Doesn't have to be of everyone. What's your best-case scenario? What does Purdue have to do to be able to demonstrate the value of our education?

DB: You don't ask easy questions! *laughter*

FD: Sorry!

DB: I don't put a lot of face value in a lot of these tests. I don't see any single test representing the diversity of degrees and programs we offer. I think if you want to look at some type of testing protocol, I think there's enough disciplines that have either licensure exams or discipline-specific protocols. If we really want to get into some type of testing

procedure, I think you have to go that route, and see how our students score on a test that represents that major, that discipline, and how are our students doing on that.

FD: Can you offer any types of information that you prefer? So, for these different types of tests-- as someone who works in assessment, what do you think is the gold standard of how to tell if a college is doing a good job? Or is it never one thing, is it a holistic approach?

DB: Yeah, I would say it's always a holistic kind of thing. You look at-- it's really looking at the alumni, and the companies that hire our students, to see what they've done with that Purdue degree. This Purdue Gallup that looks at the alumni five years out and asks-- did your college degree make a difference? Are you happy? Are you successful in your own mind? My own sister got a masters degree. She's a stay-at-home mom. She's happy as can be! So she's a success in your own mind.

FD: So life satisfaction is a big thing.

DB: Yeah. Or how you've given back to your community. Teachers don't make as much as electrical engineers make-- do we not want to graduate a lot of teachers? I think teaching is important.

FD: So just speaking for yourself, not in any kind of official capacity, how good of a job do you think Purdue is doing at educating undergraduates right now? In an unofficial sense, how do you think we do?

DB: I would say the majority of our programs do very well. We have a few that are struggling, but I think that will be true of any university. I think in general our faculty care about students, and our students have academic and non-academic success. Our students are a reflection the quality of our faculty.

FD: So do you agree that the average undergraduate who comes to Purdue and leaves with a four year degree receives a strong undergraduate education?

DB: I do.

Brooke Robertshaw Interview Transcript

3/16/2015

FD: Can you please tell me your official position in the Office of Institutional Assessment and what your defined role has been in this assessment project?

BR: So I'm a Data Analyst in the Office of Institutional Assessment. And my role in the critical thinking testing was really logistics. I'm the one who-- like all the emails came from, even though the said they came from Mitch, they were coming from me. I mean, I coordinated the whole thing. And there's a report that you haven't seen yet, I've wrote up a report reporting on things. Reporting on the pragmatics of it. And then also my findings.

FD: So then, as I understand it, there have been three separate piloting ventures that have been attached to this project, is that correct?

BR: My participation -- I don't know if there was things piloted before me -- all at the same time, so during BGR [Boiler Gold Rush] and right at the beginning of the 2014-2015 school year, three tests. Actually it was five tests: it was two surveys, so we did these two intercutural learning as well-- called the MIGUDS and the GPIs-- and then the CLA+, the C AAP, and the CTCTS. That all happened right at the same time. Our goal was to make it all happen during BGR, but that was not the case.

FD: And I have, um, so, what has been at least publicly, or has been released to me from administration is the CLA+ results from that, um, and not numerical data but discussion of the other critical thinking and the survey results that went along with that. Um, and the, the administration has chosen not to make public the results of the other tests, is that correct?

BR: I didn't even know that they had made public the CLA+.

FD: Yeah, so I've been given access to the numbers from the CLA+. So, um, without speaking on the results of those tests, is it fair to say that the purpose of doing those other tests was to try to cross-validate the CLA+?

BR: My understanding-- and I'm the lowest person on the totem pole-- my understanding is that what we were doing is we were looking to see which tests we wanted to use. So we were looking at how to, you know, do all tests... do our students... are all three of these tests measuring our students the same? So, how does that, you know, how does that look across the three tests, where are our students falling? Just sort of looking at, and then also, I mean, also on the side of which, I don't know if the higher up are interested, but the pragmatics of the tests as well. I mean, how easy is it, what's the cost? This is a university, and this is a Mitch Daniels university, and it's all about looking at, you know, what's the cost as well?

FD: It's my understanding that the CLA+ is not only the most expensive, it's the most expensive by almost twice as much as the next instrument.

BR: It is the most expensive. Have you seen the prices?

FD: I have, yeah. I think it's like $35 a test for the CLA+, and like $12.75 for the CTCTS. So was ETS's Proficiency Profile part of this?

BR: No, the CAAP was, but no the Proficiency Profile.

FD: So, I know that ETS at one point was going to come to campus and do their own piloting. I was given access to...

BR: Ah, yeah, that's the one we didn't do!

FD: Ah, OK. That's the missing chunk in my timeline. I have a proposal document where ETS says, OK we're going to come in and do this for X dollars, and so that hasn't happened?

BR: That hasn't happened. Yeah, so that.... So, sorry I'm sorry fishing around to know what you... have found out...

FD: No, I understand. [Robertshaw asks to go off the record briefly.]

BR: So that didn't happen. We were talking about doing that this spring. I think that was just a practical thing, a practicality thing.

FD: Thanks. I've been trying to track that down. People aren't always sure what they're allowed to speak about and about what.

BR: Yeah, so knowing that you knew that they were.... So before I left for Jordan, I talked to Diane, I said "If we get the go-ahead to do this, send me an email, I'm the one on the team who knows how to do this, so I'll get this going from Aman."

FD: And you decided not to do it.

BR: I don't know where that is right now. I just know that I haven't, I know that it hasn't happened, and it doesn't seem like it's going to.

FD: Right. So, to the degree that you are able to and that you feel comfortable, um, can you comment a little bit on um how you think the piloting went? And, again, not expecting you to say anything about whether or not you think the test is wise, but do you think the piloting itself served the purpose it was intended to, and do you feel like it provided sufficient information to be able to make an informed choice?

BR: Not representing the OIA is that, I don't think we accomplished what we wanted to. There was an issue with getting enough bodies. It was extraordinarily difficult. I mean, if

you look at websites, you can get this information publicly, if you Google like "Recruiting students to do the CAAP, the CTCSC, the CLA+".... they've got these whole marketing schemes! Universities develop whole marketing schemes to get students to come do this stuff. I found out like three weeks before we started doing this that I had to get these bodies. I don't think we've collected... I personally don't think we've collected enough information. If I look at it through the lens of a social science researcher, I mean for me my life, for like three weeks there, revolved around sending out emails and like, what am I gonna upload to Qualtrics today? And then send out an email, and then a reminder. So, yeah that was my life for about three weeks was, how do we get people in, and we don 't have any sort of concerted marketing team.

FD: Are you aware of what the final report is going to look like? In other words, have you guys discussed internally some kind of final report to be able to provide to the Daniels administration, to the faculty, to the community at large? In other words....

BR: There is a report that I wrote.

FD: Um, is that the one, do you know if that was the one that was shared at the faculty senate meetings?

BR: As far as I know. I would have to check with Diane to verify... the stuff that I've put together is the stuff that's been sent, that Brent sent out.

FD: Yeah, and that's been made publicly available. [interview goes off the record]... There hasn't been what I would call a great deal of secrecy about this...

BR: Oh, I think it's been wicked secret.

FD: Well, people have been forthcoming with information for me, but not a lot of coordination about who is allowed to see what, when. One thing that Brent said to me is that Purdue is a famously siloed university, and that part of the difficulty for you guys in the Assessment office has been working across the various divides, um, between different parts of the institution. Um, would you say that that's fair, that it's been difficult to coordinate?

BR: That hasn't been my experience, but I work on these various high-level projects.... Working with y'all on English 106, that's a very siloed project, right? ... Within various high-level project, yes, those are siloed. The only overlap that happens is when I'm sitting in someone's office and I'm like, hey, did you get that email. And then there's this whole critical thinking stuff that I did. Yeah, it is very siloed in some ways.

FD: A concern that Diane mentioned when I interviewed her about any of these tests is that they don't provide enough information on what particular skills students are succeeding on or struggling with. Her concern is that an individual number, and that number's place on a

distribution, does not give the institution a sufficient understanding of how to direct pedagogy. Do you feel that the CLA+ in particular provides enough information?

BR: No. There is one test that does. Not representing the Office of Institutional Assessment, just speaking for myself, it's not the CLA+.

FD: What is your preference?

BR: Did Diane share her preference?

FD: She said she has significant reservations about all of these instruments because she thinks that they are reductive.

BR: I do have a preference. It's not the CLA+. When push comes to shove, though, we have to do this. Like, I have all these epistemological... for me, my whole job is, we have these kind of epistemelogical, ontological opinions on things, but when it comes down to it if Mitch tells us to do things... we have to set that aside. As somebody who does this stuff, I think critical thinking tests are bogus. I don't think you can measure critical thinking in a test.

FD: And for me, on thing I've found more and more in this research is that there's a presumption from people that criticism of these instruments comes from a political standpoint. From my perspective, the problems aren't political. They come from the perspective of old school, hard nosed social science. In other words, their failings aren't political. Their failings are issues of sampling, issues of representation, issues of reliability and validity, rather than political.

BR: Once the report's been made fully public, and I know what's been made public, I'd be happy to chat more with you about that sort of thing. [interview goes off the record]

FD: Switching gears a bit so that you can feel freer to speak: one of the issues that's cropped up is that President Daniels is very interested in doing a longitudinal test. By default the CLA+ is a cross-sectional instrument, though it can be used longitudinally. Do you think that looking at the test longitudinally is more valid? And do you think that, at an institution like Purdue, it's practically feasible to do a true longitudinal study?

BR: I think a longitudinal study... if you look at social science research, using the same students, getting a representational sample across all of our demographics and colleges, a diverse array of programs of study... I think that that's more valid than doing this cross-sectional idea. Because then we really can... I mean I'm talking about the same students, the exact same students every year... then we can look at, we can bring in, well this student did this and this and this. You can look at your program of study. Somebody in Engineering, are their skills developing differently than somebody in Education or somebody in Agriculture? Ultimately, you'd think that students in the end when they leave are gonna get the same experience. But Engineering freshman classes-- I'm sure there are students on this

campus that aren't going to get the ENG 131/132 experiences until they reach their senior years. Or student teaching for students in the department of Education. That's a very different experience!

FD: Right.

FD: So there's research that's been done that shows that the outcomes of these kinds of tests are very susceptible to differences in student motivation. Do you think that there are any ways to control for that?

BR: Well, you can stick time on task in as a covariate, try to control for it statistically.

FD: Use time as a statistical proxy for motivation.

BR: Yeah. I get really concrete... my world has become less and less abstract since I'm in the office of assessment, I'm very concrete. Right now I'm working with R. I mean, if we used time as a covariate, then... but that's assuming that time equals motivation. How do you measure motivation? How do you know that somebody that took less time is less motivated?

FD: Especially since that NPR interview, there's been an assumption that I'm against any kind of these tests. That's not the case. But that motivation issue is huge. You can see a worst-case scenario where the administration takes the test very seriously, but the seniors don't. So we show less learning than is actually happening.

BR: Look at K-12 education. All of this testing that we do... what's happening with No Child Left Behind is that they're attaching test scores as a judgment about the teacher and about the school. I can understand that you'd attach something like the CLA+ as a judgment on the university. But I think that's....

FD: It's fair to say that you're skeptical about being able to control for motivation.

BR: Yes. How do you motivate a student? What do you do to really make sure?

FD: Right.

BR: There are proposals that they stick the score on their transcript. Well how many employers really care about that?

FD: Procedurally, practically, do you have any immediate plans for more piloting or research? Or is your role done from here?

BR: As far as I know, my role is done.... we're done piloting. It really sounds like there's a definite bias... since only one of those tests has been released to you, well... [interview goes off the record and ends.

VITA

VITA

Fredrik deBoer was born and raised in Middletown, Connecticut, where his father worked as a professor and his mother as a nurse. He has two brothers and a sister. Fredrik graduated from Middletown High School in July of 1999. He earned a Bachelor of Arts in English from Central Connecticut University in 2004, a Master of Arts in Writing and Rhetoric from the University of Rhode Island in 2011, and a Doctorate of Philosophy in Rhetoric and Composition from Purdue University in 2015. He is tired, but feeling fine.

.