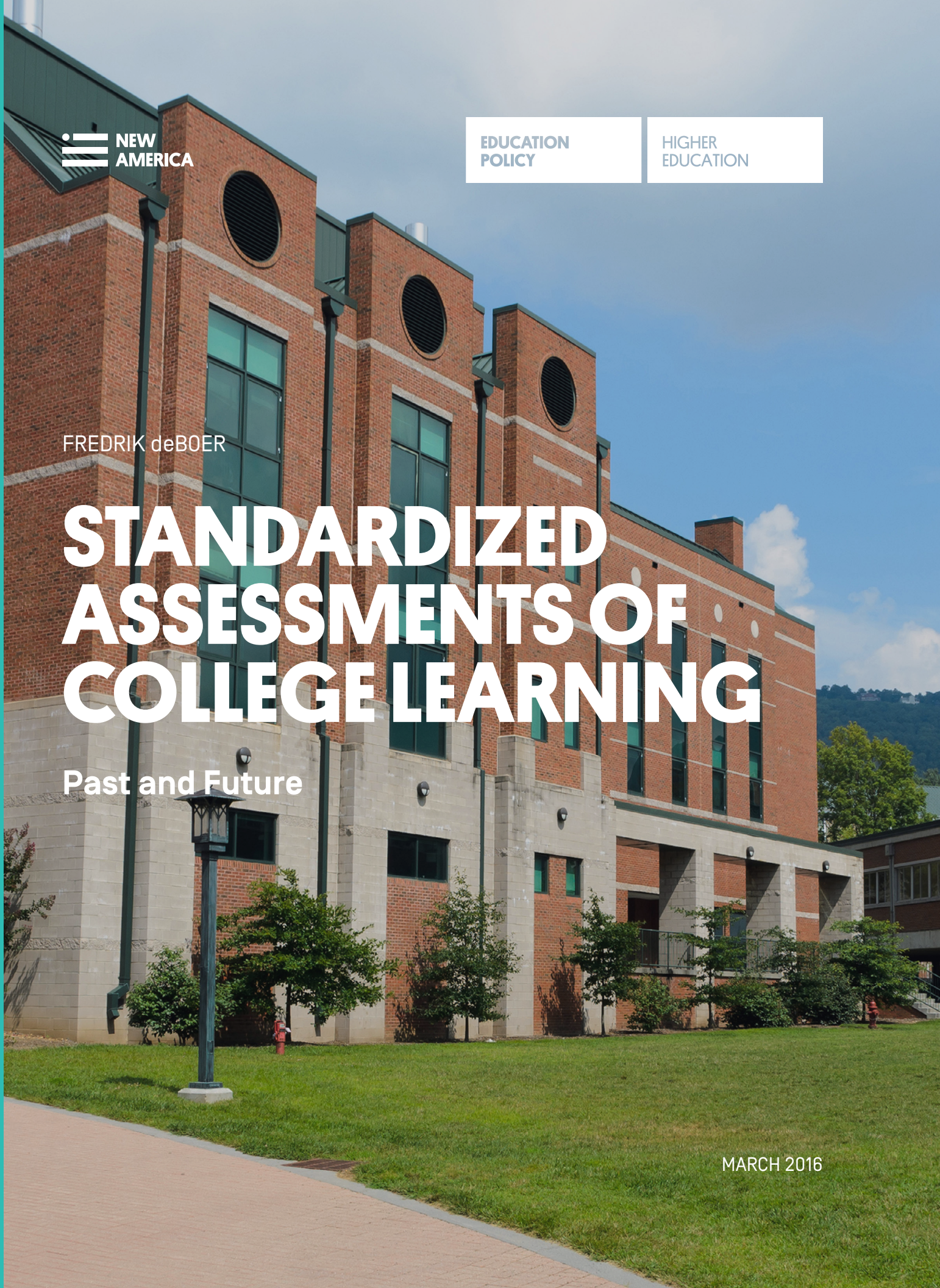

NEW AMERICA

EDUCATION POLICY | HIGHER EDUCATION

FREDRIK deBOER

STANDARDIZED ASSESSMENTS OF COLLEGE LEARNING

Past and Future

MARCH 2016

About the Authors

Dr. Fredrik deBoer is an academic and writer whose scholarly work concerns writing assessment, applied linguistics, and higher education policy. He received his PhD in Rhetoric and Composition from Purdue University in 2015. His writing has appeared in *The New York Times Magazine*, *Harper's Magazine*, *The Los Angeles Times*, *The New Republic*, *Politico*, and many others. He is currently a Limited Term Lecturer at Purdue.

Acknowledgments

We would like to thank the Lumina Foundation for its generous support of this work. The views expressed in this report are those of its author and do not necessarily represent the views of the Lumina Foundation, its officers, or employees.

About New America

New America is committed to renewing American politics, prosperity, and purpose in the Digital Age. We generate big ideas, bridge the gap between technology and policy, and curate broad public conversation. We combine the best of a policy research institute, technology laboratory, public forum, media platform, and a venture capital fund for ideas. We are a distinctive community of thinkers, writers, researchers, technologists, and community activists who believe deeply in the possibility of American renewal.

Find out more at **newamerica.org/our-story**.

About the Education Policy Program

New America's Education Policy Program uses original research and policy analysis to solve the nation's critical education problems, serving as a trusted source of objective analysis and innovative ideas for policymakers, educators, and the public at large. We combine a steadfast concern for low-income and historically disadvantaged people with a belief that better information about education can vastly improve both the policies that govern educational institutions and the quality of learning itself. Our work encompasses the full range of educational opportunities, from early learning to primary and secondary education, college, and the workforce.

Our work is made possible through generous grants from the Alliance for Early Success; the Foundation for Child Development; the Bill and Melinda Gates Foundation; the Evelyn and Walter Haas, Jr. Fund; the HeisingSimons Foundation; the William and Flora Hewlett Foundation; the Joyce Foundation; the W.K. Kellogg Foundation; the Kresge Foundation; Lumina Foundation; the McKnight Foundation; the Charles Stewart Mott Foundation; the David and Lucile Packard Foundation; the J.B. & M.K. Pritzker Family Foundation; the Smith Richardson Foundation; the W. Clement and Jessie V. Stone Foundation; and the Berkshire Taconic Community Foundation.

Find out more at **newamerica.org/education-policy**.

Contents

INTRODUCTION

The American university is currently undergoing a period of often-uncomfortable public scrutiny. Rising tuition costs and attendant higher student loan debt loads have caused both considerable human hardship and considerable criticism of individual institutions and the college education system as a whole. In the past decade, a new wave of college students has entered our institutions, many of them non-traditional students who require more support and deeper institutional investment to educate effectively. Graduation rates and retention remain areas of concern, with many more students beginning college than finishing. Meanwhile, state governments across the country continue to slash public investment in higher education.

Under these conditions, a call has risen from many corners: the call to assess. In the world of public K-12 education, the United States has seen a rapid increase in the amount of standardized testing taking place. This development has proven controversial, but there is no questioning the overall national trend of more assessment and more data collection. In the American university, however, standardized assessment remains a nascent endeavor. Unlike K-12 schools, which have long been subject to legal and infrastructural pressures that result in standardization and homogeneity, universities have traditionally been individual, self-directed institutions. Private universities, in particular, have often functioned as their own worlds, operating under idiosyncratic rules and subject to few external authorities. This lack of standardization among universities both makes it more difficult to assess college learning and harder to coordinate and standardize such assessments.

There are, additionally, legitimate concerns about the collection and analysis of assessment data on the college level. Unlike in the K-12 world, our university system lacks a history of comparative assessment of schools and programs. Although many established tests and assessment systems exist to evaluate individual students, comparative assessment of universities and their programs remains a nascent field, with fundamental questions of the validity, reliability, and fairness of such questions still largely unexplored. Many university educators have understandable fears of the consequences of assessment systems, worrying that they will result in an erosion of faculty control of curriculum and bring an end to flexible, context-specific teaching. Effective assessment of student learning in any context represents a significant challenge, and controversies persist at all levels of education about which methods of data collection and analysis are most effective and appropriate. In common with discussions of K-12 testing, some fear that the creation of widespread testing system at the college level will lead to teaching to the test and invite test fraud. Finally, large-scale assessment will undoubtedly require the expenditure of significant

Effective assessment of student learning in any context represents a significant challenge, and controversies persist at all levels of education about which methods of data collection and analysis are most effective and appropriate.

resources, at a time when colleges and universities are already under pressure to tighten their belts.

Still, the push to assess remains clear and strong, and this pressure comes from the highest offices in our government. With the college wage premium still large, the route to financial stability and a successful life for many Americans will entail receiving a quality college education. Given the costs of tuition and the potential hardship that student borrowers face, the moral need to ensure that our colleges are teaching effectively remains clear. While serious challenges to effective college assessment exist, these challenges can be met if approached with flexibility and commitment. What's more, there is no need for assessment of college learning to jeopardize the traditional ideals of faculty autonomy, institutional independence, and the pursuit of liberal ideals beyond vocational training. If undertaken with care, assessment need not unduly burden students, teachers, or administrators.

What follows is a discussion of the current political pressure to assess college learning; a discussion of some of the basic requirements for an effective collegiate learning assessment system; a brief look at extant tests to assess student learning on campus; challenges to the validity and reliability of these instruments; and a proposal for locally-controlled disciplinary assessments that could interface with standardized assessments of critical thinking skills, giving us a more full picture of student learning and including faculty in the assessment process in doing so.

THE POLICY CONTEXT AND CURRENT EXIGENCE

Like many broad changes to our system of higher education, the current assessment movement was born in the world of government policy. Universities, by design, are slowly-evolving institutions; they tend to change only when compelled to by governmental or economic pressure. The most obvious and consequential source of the current push to undertake comprehensive assessment of undergraduate education stems from initiatives of the past two presidential administrations. Both the Republican George W. Bush administration and the Democratic Barack Obama administration have called for more assessment of college learning, demonstrating the bipartisan force behind such proposals. Though the origins of the current assessment mandate stretches back much further, these executive initiatives are the most relevant to current conditions.

Arguably, the most important impetus for the current collegiate assessment movement has been the Bush administration's Commission on the Future of Higher Education, referred to as the Spellings Commission, and the report it composed. Obviously, given that the report was commissioned on September 19th, 2005 and released on September 26th, 2006, its relatively recent publication 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 U.S. 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 has endorsed a bipartisan vision for public policy and has represented the commission as non-ideological. 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."[1]

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."[2] This competitive focus persists throughout the entire document. Again and again, the exigency for improving our colleges and universities is represented as a matter of 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."[3]

The Spellings Commission called for reforms in five major areas: access, affordability, quality, accountability, and innovation. The area of most direct relevance to this paper, and which has had the most immediate policy impact—and controversy— is accountability. In particular, the finding of direct relevance to this project 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."[4] 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."[5] 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."[6]

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; 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. First, the Collegiate Learning Assessment (CLA), a prominent standardized test of college student learning. Second, the National Survey of Student Engagement and the Community College Survey of Student Engagement, a research effort of Indiana University designed to investigate educational practice at the collegiate level, such as how much time and effort students invest in learning, the number of books and papers typically assigned, and what the average requirements are for earning an American bachelor's or associate's degree. Third, 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.[7] 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 cannot help but have a powerful impact on the visibility and viability of the CLA (and its successor, 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."[8] Perhaps keeping in mind the scattered and inconsistent policy response to *A Nation at Risk*, the Reagan-era educational policy document that identified broad failures in the American educational system and called for vast reforms, 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 often deeply entrenched. 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."[9] 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."[10]

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—and a powerful lobbying organization—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," Ward argued that the report's recommendations were too formulaic and specific to address the diversity of collegiate institutions or their unique problems. This 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."[11] Although the "no" vote of a single member had little impact on the commission, the lack of unanimous consensus was something of a black eye. 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."[12]

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. 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. During the reauthorization for the Higher Education Act, Congress insisted that the legislature be given primary control of the National Advisory Committee on Institutional Quality and Integrity, which oversees college accreditors. Removing the recommended standard metrics had the unfortunate consequence, in Zemsky's telling, of shifting the information-gathering 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."[13] 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. 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 largely been a matter of continuity with that which came before it. That continuity, however, has occurred in a rapidly changing American economy.

The subprime mortgage crisis of 2009 prompted a massive spike in unemployment across the American economy, sending many workers, college educated and not, into severe financial distress. Educated workers, as they long had, continued to enjoy both a wage premium and a significantly lower unemployment rate. In 2009, the first year of Obama's presidency, Americans holding a bachelor's 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. This advantage, however, masked deep problems. To begin with, while the advantage in unemployment rate was impressive, the typical 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.[14] 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."[15] 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.[16] 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. The federal student loan system was overhauled alongside the PPACA in 2010. 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. In this title, the Obama administration referenced its Race to the Top program in K-12 education, which spurred many states to develop policies in accordance with administration policy preferences. "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."[17] 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.[18]

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. 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.[19]

The proposal calls for legislation that will ensure that "taxpayer dollars will be steered toward high-performing colleges that provide the best value."[20] 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 [21]

Though the exact formula for such rankings went unexplained (and would ultimately never be clearly delineated), this proposal represented 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 created an enforcement mechanism previously missing from past reform efforts. In its insistence on new, transparent assessments of college outcomes, the Obama proposal clearly interfaced 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. 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 standardized test to be rewarded with college credits, as part of a broader competency-based aspect of college education.

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."[22] 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."[23] "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 that "We need to give parents and students access to appropriate and robust metrics... so the overall idea is a good one."[24] Similarly, Georgetown University president John J. DeGioia expressed support, saying that "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 McGuire 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."

Like so much else about the Obama administration, then, the proposed policies regarding higher education reform have been controversial. The fierce opposition to the administration's rankings plan endured, and ultimately, that plan has been abandoned. In June of 2015, after several years of attempting to develop specific guidelines for the ranking system and appeals to stakeholders to get on board, the administration publicly distanced itself from its plan to create formal rankings. Instead, the administration argued for better consumer access to the types of information that might have gone into the creation of these rankings, such as graduation rates, real costs of attendance, post-graduation employment metrics, and similar data. Undersecretary of Education Ted Mitchell was quoted in the *Washington Post*, saying "We have decided the best way to rate colleges is to put the information and the tools in the hands of people who want to make those comparisons."[25] What remains unclear is whether parents and students possess the kind of understanding of these metrics that might make them actually useful for this purpose.

The abandonment of the rankings plan represents a major policy pivot for the Obama administration. Given that a new president will be elected in November of 2016, the future of higher education policy generally and college learning assessment specifically appears uncertain. What's more, the vociferous criticism of the proposal, and the effectiveness of university lobbying efforts to oppose the plan, may make it seem like the push to assess has run aground. Yet if we take a broader view, it should be clear that the pressure on colleges to demonstrate the effectiveness of their instruction remains powerful. Successive presidential administrations have demonstrated bipartisan interest in some form of interoperable (if not standardized) system of assessment of collegiate learning. There are few issues on which Republicans and Democrats have shown greater agreement than the expansion of educational testing. The eventual winner of the presidential campaign will go a long way towards determining the specifics of federal higher education policy moving forward, but there is little question that the issue of assessment in higher education will endure.

In order to understand how such assessments might unfold in the coming decade, it's necessary to consider the state of existing tests of college learning, their origins, and their particular advantages and drawbacks.

PRECURSORS AND ANALOGS

Though discussions of how best to teach and learn stretch back to the origins of formal education, the science of educational assessment is a relatively new field. The basic foundations of assessment theory were largely developed in the late 19th and early 20th century, during a period of democratization of education and the rise of public schooling. With a great influx of new students, most of whom lacked the economic and social privileges of those who previously had access to formal education, teachers and administrators needed new tools to guide pedagogy, and in particular to sort students into various levels of prerequisite learning and ability. At the same time, the world of cognitive and developmental psychology was undergoing rapid growth. New statistical and psychometric techniques were being developed regularly, in part to fit the needs of the massive, modern armies of the time, which increasingly required the ability to sort soldiers and officers into a hierarchy of intellectual ability. Over time, these techniques and insights filtered down into the schooling of children. These developments were largely relegated to elementary and secondary education.

In contrast, until recently there has been 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.[26]

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.[27]

> **Until recently there has been little organized development of assessments of higher education. Colleges and universities remained largely independent entities, free to dictate curricula and standards on their own.**

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. 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.[28] 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."[29] 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."[30] 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.[31] This complaint would later also be levied against tests designed to assess college learning. As with 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."[32]

> **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.**

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. Some tests attempt to address this through problem through value-added models, although these models entail controversies of their own.

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 Richard Shavelson, an expert in higher education assessment and test developer, writes in his 2009 book *Measuring college learning responsibly: Accountability in a new era*, "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."[33] Multiple choice testing and similarly reductive measures play into typical fears about educational assessment writ large: that these instruments ultimately narrow the definition of educational success into artificial and limiting constructs which cannot meaningfully demonstrate the many ways in which students learn and grow. Worse, if tests are indeed reductive in this way, the feedback loop of pressure to demonstrate student growth in assessments could result in changes to the curriculum that restrict our definition of what education can and should accomplish. The GRE and similar entrance examinations do little to address any of these concerns.

Clearly, if the higher education assessment mandate is to be fulfilled, a new measure of collegiate learning is required. In the past decade, several organizations and corporations have developed tests intended to meet this need. To what degree these tests find purchase in American colleges and universities, and which among them are widely adopted, will go a long way towards determining the future of college-level assessment.

THE COMPETITORS: MAJOR EXTANT TESTS OF COLLEGE LEARNING

Educational assessment is big business in 21st century America. Many corporations, including some of the largest education businesses in the world, have entered the educational testing market, providing tests and study materials for students and instructional materials for teachers. Nonprofit organizations have taken part in this expansion as well. What exactly it means for an organization to maintain nonprofit status has become an issue of controversy in the educational world, with many arguing that purportedly nonprofit institutions effectively function as profit-seeking entities. (Indeed, in 2009 the political advocacy group Americans for Educational Testing Reform filed suit to revoke the nonprofit status of ETS, and the group has argued that the College Board and ACT are similarly abusing their nonprofit status.) Whatever the case, a large number of deep-pocketed organizations have developed or are developing standardized tests and assorted materials. The large majority of this development occurs in the world of K-12 education, where the sheer scale of the American school system enables vast profits.[34] But precisely because standardized assessment of higher education remains so nascent and open, the field is an attractive market. Additionally, standardized tests benefit from strong network effects; the more institutions use a given test, the more broadly understood its scores become, and the more attractive the test becomes to institutions seeking test instruments. The SAT remained the dominant force in high school testing for so long in large part because everyone knows the test and can interpret its scores without much background research. Therefore, the competition to establish early inroads at prominent universities may prove especially fierce.

Given the number of tests that are currently available to colleges and universities, and this still-early stage of the development of such instruments, it's impossible to name all of the potential competitors in this paper. However, by examining the mechanisms of some of the leading tests, we can understand the particular issues that will arise in any large-scale collegiate assessment systems. For the purpose of this paper, I will restrict myself to three: the Council for Aid to Education's Collegiate Learning Assessment+ (CLA+); the Proficiency Profile, developed by ETS; and the Collegiate Assessment of Academic Proficiency (CAAP), developed by ACT, the organization behind the ACT test for high school junior and seniors. I choose these three instruments for several reasons. First, because although the number of tests taken and scored often go undisclosed, there is reason to believe that these are three of the most popular instruments measuring college learning, no doubt owing in part to the prominence of their

developers. Second, these three tests were part of a major validation study undertaken in 2009 by the organizations that develop them, and overseen by the Fund for the Improvement of Postsecondary Education (FIPSE), a subsidiary of the Department of Education that provides funding for research in college education. As previously noted, there remains little in the way of external validation of these instruments, owing to the desire of developers to maintain industry secrets and test security. The existence of FIPSE's Test Validity Study (TVS) report represents an essential benefit to examining these tests as potential wide-scale instruments for the assessment of college learning.

The Collegiate Learning Assessment+

The CLA+ is a product of the Council for Aid to Education (CAE), a New York-based nonprofit that has previously been involved in providing information on funding sources for institutions of higher education. The CLA+ is the successor to the CLA, an earlier version of the test that has been used in a great deal of collegiate research, including the famous (or notorious) *Academically Adrift*. The book, written by Richard Arum and Joseph Roksa and published in 2011, ignited a firestorm of controversy for its claims that little learning happens in college at all, with many in politics and media seizing on the research and many in academia questioning its methodology.[35] Like most such instruments, the CLA+ is intended to measure broad skills in critical thinking, quantitative reasoning, and similar cognitive abilities, rather than to assess a student's knowledge in a particular subject matter. The CLA+ is derived from a criterion sampling philosophy, meaning that it stems from a belief that the type of intellectual and academic abilities it seeks to assess cannot be meaningfully disaggregated from each other and must be measured in concert. This approach stands in contrast with the typical psychometric approach to educational testing. In that philosophy, individual aspects of student cognitive performance can be disaggregated from each other, broken down into subscores and indicators that make up different pieces of a student's ability. For example, test developers will frequently provide score reports that break down not only into broad subject areas such as math or reading, but into smaller subunits, whether concerned with subject matter like algebra or vocabulary or metacognitive skills like problem solving and making inferences. Though broad subject areas are outlined in the CLA+ rubric, the CAE maintains various aspects of college performance must be assessed in concert in order to be effectively tested and understood.

> **Educational assessment is big business in 21st century America. Many corporations, including some of the largest education businesses in the world, have entered the educational testing market, providing tests and study materials for students and instructional materials for teachers.**

The CLA+ has two major sections, the Performance Task and the Selected-Response section, which replaces Analytic Writing. The Performance Task is described by CAE:

> The Performance Task presents a real-world situation in which an issue, problem, or conflict is identified. Students are asked to assume a relevant role to address the issue, suggest a solution, or recommend a course of action based on the information provided in a document library. A full CLA+ Performance Task contains four to nine documents in the library, and students have 60 minutes to complete the task. The Document Library contains a variety of reference sources such as technical reports, data tables, newspaper articles, office memoranda, or emails. The Performance Task measures Analysis and Problem Solving, Writing Mechanics, and Writing Effectiveness.[36]

The Performance Task is graded by human raters, using a rubric that is divided into three areas of focus: 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).[37]

The Selected-Response section consists of 25 multiple choice questions, with 10 questions concerning Scientific and Quantitative Reasoning, 10 concerning Critical Reading and Evaluation, and 5 which assess the student's ability to critique an argument. The test costs $35/student, has a 90-minute time limit, and is taken on a computer. The CLA+ was the subject of my dissertation research.

The Proficiency Profile

The Proficiency Profile is developed by the Educational Testing Service, the nonprofit organization behind other major standardized tests like the SAT and GRE. The Proficiency Profile measures four major areas: critical thinking, reading, writing, and mathematics. All of these sections utilize multiple choice instruments, even the writing section, which seeks to measure this ability through tests of grammar and sentence correction. Additionally, institutions can add an optional essay testing section. The essay test is rated by an automated (computerized) rating system and delivers scores on a 1-6 scale, which is typical of such short-answer essay examinations. The test is available in a 108-question, two-hour format or an abbreviated, 36-question, 40-minute format. The test is also available in either a computerized format or a pencil-and-paper format. The test costs between $12.50 a student and $16.50 a student, depending on whether the tests ordered are the standard or abbreviated form and on how many tests are ordered at one time, with more tests purchased resulting in a lower cost. The optional essay section costs $5 a student regardless of number ordered.

The standard, two-hour version of the test features 27 questions each in critical thinking, reading skills, writing skills, and mathematics; the abbreviated, 40-minute version of the test features 9 questions of each type. ETS defines the content of these sections as follows:

Reading:

- interpret the meaning of key terms
- recognize the primary purpose of a passage
- recognize explicitly presented information
- make appropriate inferences
- recognize rhetorical devices

Writing:

- recognize the most grammatically correct revision of a clause, sentence or group of sentences
- organize units of language for coherence and rhetorical effect
- recognize and reword figurative language
- organize elements of writing into larger units of meaning

Critical Thinking:

- distinguish between rhetoric and argumentation in a piece of nonfiction prose
- recognize assumptions
- recognize the best hypothesis to account for information presented
- infer and interpret a relationship between variables
- draw valid conclusions based on information presented

Mathematics:

- recognize and interpret mathematical terms
- read and interpret tables and graphs
- evaluate formulas
- order and compare large and small numbers
- interpret ratios, proportions, and percentages
- read scientific measuring instruments
- recognize and use equivalent mathematical formulas or expressions[38]

The Proficiency Profile also adds the ability to include up to 50 questions written by the universities or programs that are purchasing the test.

Collegiate Assessment of Academic Proficiency

The CAAP is developed by ACT, the nonprofit organization that develops the ACT test for high school students interested in attending college. The CAAP is made up of six modules, each of which any individual institution can choose to implement or forego. These modules include Reading, Writing Skills, Writing Essay, Mathematics, Science, and Critical Thinking. Each module takes 40 minutes to complete. All of the modules are multiple choice, except for the Writing Essay section, which features a standard 1-6 scoring range typical of such instruments. CAAP testing costs between $13.75/student and $22/student, depending on the number of modules utilized and the amount of tests ordered.

The modules assess the following abilities and content areas, according to ACT:

Reading:

- Referring Skills
- Reasoning Skills
- Prose Fiction
- Humanities
- Social Sciences
- Natural Sciences

Writing Skills:

- Usage/Mechanics
- Punctuation
- Grammar
- Sentence structure
- Rhetorical skills
- Organization
- Strategy
- Style

Writing Essay:

- Formulating an assertion about a given issue
- Supporting that assertion with evidence appropriate to the issue, position taken, and a given audience
- Organizing and connecting major ideas
- Expressing those ideas in clear, effective language

Mathematics:

- Prealgebra
- Elementary Algebra
- Intermedia Algebra
- Coordinate Geometry
- College Algebra Trigonometry

Science:

- Data Representation
- Research Summaries
- Conflicting Viewpoints
- Understanding

- Analyzing
- Generalizing

Critical Thinking:

- Analysis of elements of an argument
- Evaluation of an argument
- Extension of an argument [39]

The CAAP permits institutions to add up to 9 locally-developed multiple choice questions to each module.

As discussed above, these three tests are far from the only potential instruments for universities looking to measure student learning. Many of them are substantially similar to the tests outlined here. In particular, tests of general college learning almost universally avoid assessing disciplinary knowledge, given the vast diversity in the content students learn in college, opting instead for measuring broader underlying cognitive and academic skills such as critical thinking. They also tend to provide similar information in terms of student performance and averages, identification of differences in outcomes for students from different broad groupings such as department, major, or year of study, and a given college's performance relative to national averages. Which test or tests are best remains largely to be seen. The nascent nature of this field makes it difficult to identify a single most effective instrument at this time. It's likely that no single test will emerge, absent some major policy decision at the federal government level, given the marketing efforts of the organizations involved and the competitive nature of the industry. It may be the case that pre-existing collections of schools, such as in athletic conferences of academic consortiums, may tend to adopt the same instrument so that they can more easily compare (and compete) with each other. At present, if I were to recommend an individual test, it would likely be the CLA+. The test's written Performance Task represents a more authentic mechanism than the pure multiple choice structures utilized by many of its competitors. While test developers who utilize multiple choice designs are correct when they argue that written responses like that of the CLA+ tend to be highly correlated with outcomes from multiple choice tests, there is still value in utilizing an instrument that is more alike the kinds of written responses that are an essential part of college academics. What's more, we may likely find that faculty members are more likely to accept a test instrument that utilizes long-form written responses rather than multiple choice questions. Finally, the CLA+'s criterion sampling approach, which acknowledges that college learning happens holistically rather than in easily-divided pieces, helps ensure that we don't draw unfair inferences about particular units within any given campus.

Before stronger recommendations can be developed, we will need the benefit of wider adoption, more research, and time.

CHALLENGES: VALIDITY AND RELIABILITY

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?"[40] Although this question is straightforward, its answers are multiple and complex, particularly in contemporary research. For decades, the simple notion of validity described above, sometimes referred to as "face 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. Convergent validity demonstrates how traits theoretically presumed to be related are actually related. Discriminant validity 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. Ultimately, validity is best thought of not as a single, specific criterion but as a vector, a multivariate aspect of tests that can be improved upon but never fully achieved. A test can be considered more or less valid but never fully or finally validated.

Reliability, in testing theory, 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. Therefore reliability is typically defined as a necessary precondition for validity.

The extant literature on the validity of existing assessments of higher education is limited, with much of it emerging from the developers of the test themselves. A document provided by CAE called "Reliability and Validity of CLA+" argues that the test has face 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 document argues that these responses demonstrate face validity for the test and that "it appears that we are measuring what we purport to measure on the CLA+ tasks."[41] 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.

Reliability, in testing theory, refers to a test's consistency: does the test measure different people in different contexts at different times in the same way?

For testing instruments like those considered in this paper, 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.[42] 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."[43] 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. 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 of the CLA, 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."[44] 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 and inherent reliability limits in human rating.

> **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.**

ETS provides several documents concerning validity and reliability of the Proficiency Profile. A text concerned with the validity of the Proficiency Profile, titled "Validity of the Measure of Academic Proficiency and Progress,"[45] concerns itself with construct validity, generally defined as the degree that a test represents the underlying knowledge or ability a given test is intended to measure. The document argues that the construct validity of the Proficiency Profile was established in earlier testing of its predecessor, the Academic Profile, particularly in research conducted in 1990 and 1995.[46] Additionally, the document suggests that the Proficiency Profile benefits from demonstrable criterion validity, as earlier research demonstrates an association between scores on the instrument and grade point average, class level, and amount of core curriculum completed. A 2008 publication of ETS's research arm, "Measuring Learning Outcomes in Higher Education Using the Measure of Academic Proficiency and Progress (MAPP)," reported that internal reliability measures of the four major sections of the Proficiency Profile range from .80 to .89, suggesting strong internal reliability.[47]

ACT provides a *Technical Handbook* for the CAAP that provides validity and reliability evidence. The *Handbook* references two major elements of validity, content validity and criterion validity. Content validity refers to the degree to which a test measures content considered appropriate and necessary for students in the given domain being tested. For a test like the CAAP, therefore, a test has content validity if it adequately measures the content considered essential for the college students it is intended to test. Content validity is inherently subjective and contextual and depends a great deal on a given definition of the tested construct, so ACT's opinion on the CAAP's content validity, while reasonable, does little to inform outside stakeholders about its overall validity. Criterion validity, on the other hand, involves the relationship between a test and external criteria that can be assumed to be associated with the test's constructs, and can be evaluated empirically. The *Handbook* notes that a study of 787 sophomores found positive correlations between English GPA and CAAP Writing Skills (.37), mathematics GPA and CAAP Mathematics (.34), and overall sophomore GPA and CAAP Writing Skills (median r = .36), Mathematics (.35), Reading (.38), and Critical Thinking (.34).[48] Given the inherent noise of GPAs, these correlations are adequate, if not impressive. Similar correlations were found when predicting junior year grades, for junior English GPA and CAAP Critical Thinking score (.32), Writing Skills score (.25), and Reading score (.25). Junior mathematics GPA was similarly correlated with CAAP Mathematics score (.23).

The primary CAAP reliability evidence presented by ACT concerns internal consistency. Internal consistency generally refers to a method of establishing reliability in which researchers ascertain how well different items on a test that are intended to measure the same construct produce similar results. That is, a test is considered to be internally consistent in measuring mathematical

reasoning, for example, if individual students perform about as well across different items meant to measure mathematical reasoning, relative to the intended difficulty of that item. In keeping with the industry standard, ACT uses the Kuder-Richardson Formula 20 (K-R 20) process to ascertain internal consistency. Across the various modules of the test, reliability estimates range from .92 to .84, suggesting high reliability ratings that are well in keeping with industry standards.[49]

> **This issue of motivation is a particularly acute problem for value-added metrics, as students who apply greater effort to one test administration than they do to another would artificially distort the amount of demonstrated learning.**

The previously-mentioned FIPSE study represents one of the most important documents regarding the validity of these exams. It is important to note again that the study was in fact authored by employees of CAE, ACT, and ETS. For this reason, it cannot be considered truly independent research, and outside validation of these metrics remains an essential task for establishing their validity and reliability. Still, 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.[50] 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 the desire of developers to differentiate their own tests. The study also established internal consistencies for all 13 tested modules and sections, with a mean reliability of .87, suggesting strong internal consistency across the evaluated instruments.[51]

Despite the positive outcomes from the FIPSE study, significant concerns for these types of instruments persist. 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, as students who apply greater effort to one test administration than they do to another would artificially distort the amount of demonstrated learning.[52] At present, standardized tests of college learning are low stakes tests 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 these instruments. Frequently, schools have to provide incentives for students to take the tests 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 CLA+, for one, acknowledge this problem, 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."[53]

Measuring motivation, however, is empirically difficult. Self-reported motivation scales are subject to the typical reliability challenges of all self-reported data and of Likert-type measurement scales. Researchers sometimes attempt to address these issues by measuring more objective variables that might be reasonably associated with motivation. One attempt was made at Central

Connecticut State University, which at the time utilized the CLA, the precursor to the CLA+. Dr. Brandon Hosch attempted to measure student motivation by measuring time on task, 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."[54]

Hosch also gave students a survey to report their level of motivation. While self-reported data must be taken with a grain of salt, as noted previously, Hosch found that only 34% of freshman agreed or strongly agreed that they were highly motivated on the CLA.[55] 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.[56] 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."[57]

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. 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."[58] That effect size is quite large, indicating that student motivation is a major aspect of such performance metrics, and a major potential confound. Liu, Bridgeman, and Adler suggest that this phenomenon could be expected in any test of college learning that is considered low stakes.[59] The results of this research were important enough that Council for Aid to Education President Roger 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."[60] Clearly, the impact of student motivation on standardized tests of college learning will have to be monitored in the future and represents a significant challenge to the validity of such instruments.

CRITICISM AND CONCERN

Unsurprisingly, given the stakes involved, the resources required to implement such testing, and the complexity of developing a fair and effective assessment system, these tests are controversial. Students, faculty, and administrators at various colleges have legitimate concerns with the implementation of these tests, the interpretation of their results, and the consequences of their findings.

One of the most common frustrations voiced by faculty regarding standardized tests of college learning concerns the lack of disciplinary assessment within their systems. Although there are many tests available for measuring disciplinary knowledge for various majors, none of the currently-available tests of general college learning attempts to directly assess such knowledge. That is, tests like the CLA+, Proficiency Profile, and CAAP do not attempt to ascertain how much history has been learned by History majors, how skilled Computer Science majors are at programming, how deeply Psychology majors have absorbed current theories of the science of the mind, etc. Measures such as Mathematics on the Proficiency Profile or Science on the CAAP are indicators of broad understanding of certain disciplines, but no test developers claim that these instruments effectively measure the knowledge and skills students acquire in their majors. This is by design; the tests are intended to assess students from all kinds of majors, and no individual instruments can possibly measure disciplinary knowledge from the broad sweep of subjects studied in the American university system. But this leads to understandable frustration on the part of faculty and the programs they run: how can we adequately understand student learning if we do not measure disciplinary knowledge? A potential solution to this issue is outlined below.

Another concern for faculty lies in the notion that these tests are inherently reductive and fail to accurately reflect the deeper learning and shared values of the university. Traditionally, higher education has not been intended merely to inculcate knowledge in students or contribute to their vocational skills, but also to develop within them the less tangible concepts of the liberal arts, such as emotional intelligence, aesthetic appreciation, and ethical reasoning. These values are considered particularly important in the humanities and social sciences. While there is no particular reason why such values could not be respected and valued alongside the more quantifiable skills that these instruments measure, there are legitimate concerns that those facets of college student growth that are not measured will become marginalized. It has often been alleged, after all, that subjects such as art, music, and gym have suffered in K-12 education given their lack of presence in standardized testing.

Finally, faculty are frequently concerned that assessments such as these may undermine

their control of curriculum. Faculty control over teaching and learning has traditionally been a treasured value within the higher education system. Individual departments are subject to institutional curricular mandates, and public universities are subject to guidelines from state, federal, and accreditor guidelines, of course. But to a large degree, individual departments have broad latitude to determine for themselves the knowledge, skills, and competencies that their majors should seek to establish in undergraduates. This control could be threatened if standardized assessments result in institutional pressures on individual departments or majors to raise particular test scores or subscores. This is the much-discussed "teaching to the test" effect—the fear that standardized testing will result in education that serves the needs of standardized testing, rather than the other way around.

Concerns such as these have already caused campus tensions about assessment to come to a head. An indicative example can be found in the implementation of the CLA+ at Purdue University, my doctoral institution. The administration of Purdue President Mitch Daniels attempted to implement the CLA+ to large numbers of incoming freshman and outgoing seniors, in order to demonstrate learning in the undergraduate population. Faculty, however, objected to the proposed change, arguing that the test had been insufficiently validated, that the plan for interpreting results and using them to direct administrative changes was unclear, and that insufficient infrastructure had been developed to effectively introduce the test to Purdue. The conflict between the Purdue faculty senate and the Daniels administration would eventually be described in the local media as a "Clash of Wills."[61] After more than a year of debate and negotiations, the faculty senate and Purdue higher administration eventually found a way forward, in part by agreeing to a dramatically smaller sample than originally proposed. (This reduction stemmed not only from faculty concerns but also from practical difficulties in assembling an adequate sample of undergraduate students.) The fight at Purdue, however, demonstrates the kind of conflict that will likely be a part of wide scale adoption of these instruments.

Conflicts of this type are likely inevitable, and perhaps necessary. While conflict is never pleasant, these debates demonstrate the investment that various stakeholders in the university system have in their vision of the purpose of the contemporary university and what is best for individual institutions. The interplay between faculty and administration helps to ensure that the various needs and commitments of colleges and universities are defended appropriately. Ultimately, students, parents, professors, administrators, government officials, and taxpayers all have a role to play in the oversight that should attend these major changes to the university system. One productive way to help prevent these rifts in the future lies in allowing faculty and departments to take an active role in the assessment process, as described below.

LOCAL DISCIPLINARY ASSESSMENT

As we've seen, local control vs. standardization and critical thinking measures vs. disciplinary knowledge represent two of the central tensions in the assessment of college learning. Both of these issues are matters of potential controversy in the assessment process, and for obvious reasons; they strike at the heart of faculty control of undergraduate learning and what we value in college education. Navigating these tensions will always require sensitivity and care on the part of administrators, and necessarily involves addressing institution- and department-specific concerns. If this care is applied, faculty and administrators alike can feel invested in the assessment process, and help to make assessment fair, valid, and reliable. As Edward White, Norbert Elliot, and Irvin Peckham write in *Very Like A Whale: The Assessment of Writing Programs* (2015), "How to establish the importance of both broad-based consensus and local relevance…? Such alignment begins with recognizing the importance of both broad outcomes (broad and general results of an expansive curriculum) and locally developed competencies (distinct and varied within institutions and among specific programs). Once outcomes are distinguished from more narrowly defined competencies, it is possible to reach consensus—subject to refinement and revision—across institutions."[62]

The lack of disciplinary assessment cannot be filled through any individual test. Instead, it should be ameliorated through a large variety of individual, locally-control disciplinary assessments that individual departments choose for their undergraduate students. The potential types of local, discipline-specific assessment to utilize are vast. The world of disciplinary assessment theory, research, and instruments has exploded in recent years. A 2010 article in *Assessment Update* by Theresa Ford documented this growth, listing recent books like *Assessment in Political Science* and *Supporting Assessment in Undergraduate Mathematics*, as well as conferences and organizations that concern disciplinary assessment. Various academic and professional organizations have begun, in the past several decades, to develop best practices and procedures for how to effectively measure student learning gains within their given subjects. For example, the American Historical Association has produced guidelines for curriculum and assessment in history, and has also provided input to the development of K-12 curriculum and testing such as that in the College Board's Advanced Placement U.S. History. This kind of work can help to create continuity between K-12 and collegiate education, and ensure that assessment becomes an organic aspect of classroom practice, rather than a tacked-on distraction from learning. Perhaps

drawing from these disciplinary texts, many departments will choose to develop their own local assessment procedures themselves. The ultimate instruments developed across the thousands of universities and hundreds of thousands of majors within the United States higher education system are, obviously, too potentially varied to predict. But discussing a particular assessment system for a particular discipline may help to define the shape of potential disciplinary instruments.

College writing programs are among the most universal in the higher education system, with introductory writing courses perhaps the single most commonly taken course in American college education today. This breadth is understandable, given that effective writing stands as one of the most essential underlying skills for college success, with the ability to write a paper important to almost any college student's career. The field of writing studies has developed a great deal of research concerning effective, authentic, and fair assessment of student writing ability in the past several decades. A given writing program could utilize this research to develop an internal disciplinary assessment of their classes and students that could then interface with broader, standardized tests like those described in this document.

A typical assessment system for a writing program would likely consist of the collection of student texts to be rated by trained raters. Testing every student would be inefficient and unnecessary; a reasonable sample could be derived from the overall program population. In order to achieve generalizability, students would have to be randomly selected, which can be achieved with randomization procedures that are widely available through spreadsheet software. Assessment administrators should also ensure that the randomly-selected sample contains a representative cross-section of the overall undergraduate population, stratifying the sample on racial, gender, and similar lines.

Once a representative sample has been developed, an appropriate assessment instrument can be implemented. Typically, this instrument would entail having students write a short essay response to a given prompt or a prompt drawn from a small number of possibilities. These essays would then be rated by a team of raters, trained by the programs utilizing rubrics internally developed for this purpose, with each essay rated by several raters to ensure the reliability of the ratings. Alternatively, there are several extant rubrics and essay prompts that have been developed by educational organizations and are available for broad use. Some programs might choose to apply a test-retest model, collecting essays from students at the beginning of the semester and the end, in order to ascertain how much students improve in their essay writing in that time span. Results could be analyzed statistically, to ensure that students from various subpopulations within the college community were all learning. Various aspects of validity and reliability could be tested as well. An alternative that many programs might adopt would be the collection of portfolios of student writing, where several pieces of writing are collected for each tested student and assigned a holistic score. These mechanisms are viewed by many writing instructors as more authentic and fair, but they also suffer from lower reliability and require more resources than other options. Ultimately, a given college or university would ideally assemble adequate assessment data from a variety of majors and programs, which could then be compared to the outcomes of standardized tests. Such assessments could be undertaken perhaps every three to five years. In this way, a more holistic, multifaceted, and complete picture of learning could be assembled, in a way that made faculty members intimate partners, rather than skeptical outsiders.

The time, money, attention, and resources required for this type of multifaceted assessment regime would be considerable.

RECOMMENDATIONS

I have several recommendations for the next stage of development in assessment of American higher education.

1. Standardized tests of collegiate learning must be subject to external validation. While extant standardized tests of college learning have several admirable qualities, they remain largely unvetted by external researchers. This stems in part from the fact that these tests are still rather new developments, but also from a persistent facet of the standardized testing industry: a generally closed-off nature and lack of transparency. Many researchers in education, psychometrics, and language testing regard the difficulty of obtaining research materials from testing companies as a fact of life. While testing companies will sometimes provide data, this data typically comes with restrictive stipulations for its use and for publishing the results of research performed with it. It's not uncommon for these stipulations to involve a long, multi-stage approval process. Given the importance of publishing research reports for an academic career, and given the fact that researchers feel time pressures for hiring and tenure, these delays can often act as a strong disincentive for attempting to access such data.

There are reasonable arguments for why testing companies might be stingy with their data. Maintaining test security is an essential goal of test development, and there are some legitimate corporate interests in maintaining trade secrets. But ultimately, these test developers are attempting to take on a crucial role in public institutes of higher learning, and to do so through the expenditure of public funds. They must understand that only truly independent verification of their claims to validity, reliability, and fairness can result in real public confidence. After all, the very purpose of assessment is to provide external, independent verification of educational claims, which we take to be necessary given the importance and expense involved in college education. Those offering to perform such assessment should recognize that the need for external validation falls on them, too. Researchers must vet these instruments to determine how well they work, and what the potential unforeseen consequences are of these types of assessments, for the good of all involved.

2. Faculty and local administration must be welcomed into the assessment process. I have already advocated turning disciplinary assessment over to faculty and their departments, for the reasons outlined above. This is one example of a broader necessity: ensuring that the faculty, staff, and administration of universities feel that they have a hand in directing the assessment process. Too often, assessment is posed as antagonistic to college teachers and programs. If assessment is presented in terms that seem to confirm the fear

of unaccountable outsiders undermining faculty control of curriculum, mistrust from instructors is a certainty. But such mistrust is not inevitable. If faculty and departments are reassured that they have influence over key decisions within a comprehensive assessment plan, they are far more likely to see themselves as partners in the endeavor rather than as targets. This in turn will make the implementation of such a plan far more pleasant and straightforward. The specific ways in which faculty can be brought into the fold will vary from campus to campus, and depend largely on the given power-sharing structure at particular institutions, but some effort to invite teaching staff into the assessment process is essential for creating effective relationships and lasting assessment programs.

3. Assessment of college learning should take advantage of the power of representative sampling and inferential statistics. One of the common complaints in K-12 testing lies in the time and resource commitment. Parents and teachers complain that testing occupies far too much class time, disrupting typical educational schedules and leaving insufficient time for essential learning. What's more, the resource costs of such testing are often considerable, calling into question whether such expenditure is an appropriate use of public funds, particularly given how much of it ends up in the hands of private enterprise. Both of these concerns stem from the fact that, in many contexts, K-12 testing utilizes a census approach, with almost all students taking part in the tests. The wisdom of that approach lies outside of the scope of this paper. In college assessment, however, we should take full advantage of the affordances of appropriately stratified samples and inferential statistics. With the precision and sophistication of contemporary statistics, we have no need to test all of the students all of the time. Instead, appropriate samples should be developed through consultation with statisticians, taking care that these samples adequately reflect the various forms of diversity on a given campus and can be responsibly used to draw inferences about the campus population as a whole. With appropriate sampling, monetary and time expenditures can be reduced, and assessment can take place with minimal disruptions to day-to-day university life.

4. Standardized assessments and localized disciplinary assessments should be used in concert with student outcomes data to better understand both individual colleges and our system as a whole. By outcomes data, I mean rigorously collected financial and life-satisfaction figures from college graduates. This data can potentially be collected at three, five, or ten years after graduation, or similar. This data will help us to understand if college is actually improving long-term life outcomes for our graduates. We must be careful not to slip into economic reductionism in this effort; the purpose of higher education is not merely to train workers but to educate citizens, after all, and many college graduates make life and career choices that reduce their incomes but improve their happiness. In order to present a full-fledged picture of real outcomes, we should endeavor to gather self-reported data about life satisfaction and satisfaction with one's education. This work has already begun in instruments like the Gallup-Purdue Index, a large collaborative study that examines a broad range of self-reported outcomes data from graduates of many different universities. Larger scale (though less granular) is the Department of Education's College Scorecard. Though still nascent, this effort will in time provide massive amounts of data on the economic outcomes of students from different colleges. This is a positive development overall, but such data is inherently noisy, and must be approached with caution. Differences in incoming ability, demographic factors, and the inherent variability within large-scale human data could lead to erroneous causal arguments about average earnings and the quality of a given institution. Assessment data can be used to validate and support (or, alternatively, undermine) such claims. As always, our intent must be to collect many different kinds of data and use it to draw a broad overall picture of our institutions and our system, in order to avoid falling victim to problems with any individual data source.

5. Assessments cannot be no stakes, but neither should they be high stakes. Assessing college learning will be a major undertaking. It will involve significant expense, many hours of work, and considerable effort, to say nothing of the inevitable unhappiness and growing pains that any major institutional change is likely to engender. In order to be worth it, assessment must be used to actually improve our institutions. If we take care to gather several different kinds of evidence, investing appropriate skepticism in our instruments but also taking their findings seriously, we can use assessment data to guide pedagogical and administrative practice, identifying areas of strength, areas of need, and areas where different groups of students are seeing unequal outcomes. Most importantly, we can develop a better picture of how well public resources and tuition dollars are being used to educate our people.

But we must be careful, fair, and realistic with the consequences of our assessments. It would not be fair, or even practically possible, to use contemporary assessment systems to evaluate the instruction quality of individual instructors. Even attempting to disaggregate the performance of individual programs and departments from broader institutions is fraught with difficulty. There are reasons of labor rights, faculty independence, tenure, and institutional separation of powers for this, but more importantly, reasons of basic responsible empiricism. These instruments are in their infancy, and considerable room for development and improvement exists. The vast differences in incoming ability levels of undergraduate students makes fair comparisons challenging, not only between institutions but within institutions. What's more, college students learn and grow in a large variety of settings within their institutions. They take classes not only in their majors and minors but within general election courses and electives. They learn not only in the classroom but on internships, in extracurricular activities, and during study abroad. Effectively determining which specific classes or instructions within this broad experience contribute significantly to overall learning would represent an immense empirical challenge.

None of this means that we should be nihilistic about assessment. In particular, we should feel confident that comprehensive, responsibly-implemented assessments can function well on the institutional level, helping us to understand which schools are graduating students who have made demonstrable learning gains. We have to maintain appropriate care in interpreting results, especially given the continuing controversy over value-added models and other attempts to address differences in incoming ability. But there is a great deal of useful information to be gathered. That's the spirit with which we should undertake our assessments: as a mutual collaboration between various partners in the higher education system, designed to gather and share information of relevance to institutions, students, parents, politicians, and citizens.

CONCLUSION

The future of assessment of higher education in America remains cloudy. The task of assembling and interpreting adequate data is considerable, particularly at the scale necessary to truly understand the amount of learning occurring at college campuses. Resistance and criticism, both fair and unfair, have attended every effort to create such large-scale systems in the past. The question of educational assessment is inherently political, with many actors involved, some of them concerned primarily with the profit motive. Colleges and universities represent a powerful lobbying interest in American politics, and have demonstrated a powerful ability to resist outside accountability in the past. Perhaps most importantly, we have much evidence to believe that our higher education system is succeeding in many of its core functions, and we must proceed in our pursuit of college learning assessment in a way that does minimal harm to these functioning systems.

And yet despite these considerable challenges, there are perhaps reasons for optimism. The need to assess is clear: in a world in which a college education has such economic and social power, we have a fundamental responsibility to ensure that our students are learning. What's more, with so many challenges to traditional universities, particularly in the form of online and for-profit schools, the need to demonstrate our value is greater than ever. The great challenge of effective assessment can be met, but only if the many stakeholders within the university work together to find fair, authentic, valid, reliable, and ethical systems. The controversies and debates about how to undertake this work will continue. But with open dialogue and an attempt at mutual understanding, fair and effective assessment of college learning is possible. It's time to get to work.

Notes

[1] Spellings, Margaret. *A test of leadership: Charting the future of US higher education*. U.S. Department of Education, 2006, 33.

[2] Spellings, *x*.

[3] Ibid., *ix*.

[4] Ibid., *vii*

[5] Ibid.

[6] Ibid., 24.

[7] Ibid., 22.

[8] Ibid., 23.

[9] Ibid., 24.

[10] Ibid.

[11] Lederman, Doug. "18 Yesses, One Major No." *Inside Higher Ed*, 11 August 2006.

[12] "AAUP Statement on Spellings Commission Report," *AAUP,* 2007.

[13] Zemsky, Robert. "The unwitting damage done by the Spellings Commission." *Chronical of Higher Education*, 18 September 2011.

[14] Weissmann, Jordan. "How Bad is the Job Market for the Class of 2014?" *Slate.com*, 8 April 2014.

[15] "Student Debt and the Class of 2012." *Project On Student Debt*. Institute for College Access and Success, December 2013

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

[17] "Remarks by the President on College Affordability, Ann Arbor, Michigan." *WhiteHouse.gov*. Office of the Press Secretary, 27 January 2012.

[18] Ibid.

[19] "FACT SHEET on the President's Plan to Make College More Affordable: A Better Bargain for the Middle Class," *WhiteHouse.gov*. Office of the Press Secretary, August 22, 2013.

[20] Ibid.

[21] Ibid.

[22] Selingo, Jeff. "President Sees an Obamacare Solution to Higher Ed's Problems." *Chronicle of Higher Education*. Chronicle of Higher Education, 22 August 2013

[23] Stratford, Michael. "Margaret Spellings on Obama Plan." *Inside Higher Ed*, 5 September 2013.

[24] Anderson, Nick. "College Presidents on Obama's Rating Plan." *Washington Post*, 13 August 2013.

[25] Anderson, Nick. "Obama administration retreats from federal college rating plan," *Washington Post*, 25 June 2015.

[26] "E-Update: The GRE Revised General Test." *Educational Testing Service*, May 2010.

[27] Shavelson, Richard. *Measuring college learning responsibly: Accountability in a new era*. Stanford University Press, 2009, 29.

[28] See, for example, Kaplan, Robert, and Dennis Saccuzzo. *Psychological testing: Principles, applications, and issues*. Cengage Learning, 2012, and

[29] 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.

[30] "What is the Value of the Graduate Record Examinations?" *Educational Testing Service*, October 2008.

[31] Hunter, John E., and Ronda F. Hunter. "Validity and utility of alternative predictors of job performance." *Psychological bulletin* 96.1, 1984, 72.

[32] Cross, Theodore, and Robert Bruce Slater. "Special report: Why the end of affirmative action would exclude all but a very few blacks from America's leading universities and graduate schools." *Journal of Blacks in Higher Education*,1997.

[33] Shavelson, 30.

[34] "The Testing Industry's Big Four," *Frontline*, 2001, http://www.pbs.org/wgbh/pages/frontline/shows/schools/testing/companies.html

[35] A follow-up study performed by the CAE, also using the CLA but dramatically expanding the sample size and increasing the time between test administrations, found significant learning gains at all tested colleges and a robust average institutional improvement. This report, titled "Does College Matter?," is available on the CLA website. The debate over *Academically Adrift* continues to this day.

[36] "CLA+ Sample Tasks." Council for Aid to Education, nd.

[37] "CLA+ Rubric." *Council for Aid to Education*, nd.

[38] ETS Proficiency Profile Content," *Educational Testing Service,* nd.

[39] "CAAP Test Modules," *ACT,* nd.

[40] Fulcher, Glenn. *Practical language testing*. Routledge, 2013, 19.

[41] "Reliability and Validity of the CLA+," *The Council for Aid to Education*, 2012, 5.

[42] Ibid., 3.

[43] Ibid.

[44] Ibid.

[45] Measure of Academic Proficiency and Progress, or MAPP, was the previous name of the Proficiency Profile. The test itself is identical.

[46] Young, John W. "Validity of the measure of academic proficiency and progress (MAPP)," *Educational Testing Service,* 2007, 4.

[47] Liu, Ou Lydia. "Measuring learning outcomes in higher education using the Measure of Academic Proficiency and Progress (MAPP)." *ETS Research Report Series,* 2008.

[48] "CAAP Technical Manual," *ACT,* nd, 37.

[49] Ibid., 18.

[50] Klein, Stephen, Ou Lydia Liu, James Sconing, Roger Bolus, Brent Bridgeman, Heather Kugelmass, Alexander Nemeth, Steven Robbins, and Jeffrey Steedle. "Test Validity Study (TVS) Report.," Fund for the Improvement of Postsecondary Education, 2009, 24.

[51] Ibid., 4.

[52] Though there is little empirical verification for this supposition at present, it's commonly conjectured that exiting seniors would be less likely to invest their highest possible effort in the test, given the distractions of graduation and the fact that they will have recently undergone the rigors of completing their degrees.

[53] "Validity and Reliability."

[54] 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, 7.

[55] Ibid., 8.

[56] "Validity and Reliability," 4.

[57] Hosch, 9.

[58] Liu, Ou Lydia, Brent Bridgeman, and Rachel M. Adler. "Measuring learning outcomes in higher education motivation matters." *Educational Researcher* 41.9, 2012, 356.

[59] Ibid., 359.

[60] Jaschik, Scott. "Tests With and Without Motivation." *Inside Higher Ed*, 2 January 2013.

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

[62] White, Edward M., Norbert Elliot, and Irvin Peckham. *Very Like a Whale*. Utah State University Press, 2015, 20.