# SYNOPSIS PAPER

Charles T. Clotfelter

September 2012

HCM STRATEGISTS

In 2011 the "Context for Success" project, sponsored by the Bill & Melinda Gates Foundation, brought together a group of scholars and policymakers to consider issues related to the evaluation of postsecondary institutions. This synopsis paper describes the project, presents an overview of the papers commissioned for the project and offers recommendations that are drawn from the working group's consultative efforts. In this overview, I begin by summarizing the rationale behind the Context for Success project and describing the historical "context" for it, noting the interest in data-driven performance policies and the success of such policies. I then turn to the studies commissioned for the project, discussing both their common themes and their differences. I pay particular attention to the outcome measures they use, the methods they employ to correct for inputs and the difference such adjustments make. I conclude by summarizing the lessons to be gleaned from this project.

## I. THE CONTEXT FOR SUCCESS PROJECT

Few would disagree that the United States needs to improve the quality and expand the reach of its system of postsecondary education. Doing so will strengthen the skills of American workers, making the country more competitive and improving the life prospects of its citizens. If postsecondary education in the United States is to be improved, it will make sense to put our scarce resources into the institutions that are most effective and to encourage all institutions to get better. To do this, policymakers will need to be able to identify which institutions are most effective in educating students. And in order to make that possible, two practical problems have to be solved. The first is how to devise sensible measures of student learning and other desired educational outcomes. The second is how to correct for the fact that students enter college at different starting points.

The second of these problems is the primary focus of the Context for Success project. Well before the project was undertaken, in 2011, there had developed a rich national debate about policies and approaches that would allow decision makers to use data on educational outcomes to identify those institutions most deserving of support. For example, an institution's graduation rate could be used as a measure of its ability to produce successfully trained graduates, ready to enter the labor market or pursue further training. Congress in effect endorsed this indicator of quality when, in the 1990 Student Right to Know Act, it mandated that postsecondary institutions report graduation rates.[1] But one nagging worry about using a

[1] The Student Right to Know Act, also known as the "Student Right-to-Know and Campus Security Act" (P.L. 101-542), was passed by Congress November 9, 1990. Title I, Section 103, requires institutions eligible for Title IV funding to calculate completion or graduation rates of certificate- or degree-seeking, full-time students entering that institution, and to disclose these rates to all students and prospective students. http://nces.ed.gov/lpeds/glossary/index.asp?id=625, 12/27/1.

measure like this is that some institutions enroll students with much stronger educational backgrounds than others, giving those institutions a built-in advantage in achieving high graduation rates that might have little to do with their own effectiveness in educating students. The difficulty is exactly analogous to that of determining the fastest runner on a track team. If runners start a race at different spots on the track, the runner who crosses the finish line first will not necessarily be the fastest. To determine who is fastest, it is necessary to correct for those starting positions. In the same way, if graduation rates are to be used to compare the educational effectiveness of postsecondary institutions, it is necessary to correct for differences in student preparedness, or differences in educational "inputs."

Another way to understand the logic of adjusting for inputs is to consider the economic concept of "value added," the increase in a product's value as it proceeds through each stage of the production process from raw materials to finished product. The baker's value added, for example, is the difference between the value of the bread and the cost of the flour, labor and other inputs to production. In the same way, a year of school's value added is the additional learning a student gained over that year, not the student's stock of knowledge at year's end. If periodic measures of knowledge were available, a straightforward approach to measuring value added would be to calculate the gain over time. This approach is widely used in K-12 evaluation schemes, owing to the availability of annual achievement tests in numerous grade levels for many students. Since this sort of information is rarely available for college students, other ways of correcting for inputs have to be devised, leading back to the track analogy.

This issue of adjusting for inputs is not merely a fine point of academic discussion. As esoteric as it might at first appear, the issue of input correction promises to be of great significance in the larger policy discussion about how best to evaluate postsecondary institutions. In terms of practical implications for actual policies, it is no longer a question of whether data will be used to measure outcomes. This is being done now. For more than a decade, a number of states have used performance measures in allocating resources.[2] Rather, the pressing question is whether adjustments should be made for differences in student inputs, and how that ought to be done. Logic surely suggests that outcome measures must be corrected for differences in inputs, as suggested by the track team analogy. But, as the discussion that took place among participants in this project revealed, the question is decidedly less easy to answer once real-world complications raise their heads. In particular, input adjustments raise two very real problems. First, such adjustments typically require

---

The graduation rate mandated by the act is the percentage of first-time, full-time students who obtain a degree from that institution within 150 percent of the "normal" completion time. So, for four-year institutions, it is the percentage of such students who obtain a bachelor's degree within six years at the same institution where they matriculated.

[2] A study showed that 21 states in the 1990s employed some performance measures in allocating resources and virtually all states monitored performance indicators as part of budget reviews. National Commission on Accountability in Higher Education (2005), p. 31.

*Context for Success is a research and practice improvement project designed to advance the best academic thinking on postsecondary institutional outcome measures. The project was organized by HCM Strategists LLC with support from the Bill & Melinda Gates Foundation. The papers may not represent the opinions of all project participants. Readers are encouraged to consult the project website at: www.hcmstrategists.com/contextforsuccess.*

more information than is usually collected, and that data collection is expensive. Second, input adjustments invariably require adding complications to what would otherwise be fairly easy comparisons, replacing transparency with opaqueness, and in the process raising suspicions and jeopardizing the political support of important constituencies. In light of these tradeoffs, we are left with no choice but to compare second-best alternatives.

The objective of the Context for Success project was to ask scholars of higher education to weigh in on the issues—both theoretical and practical—that need to be considered in designing "input-adjusted metrics" for judging the effectiveness of postsecondary institutions. With the support of the Bill & Melinda Gates Foundation, the consulting firm HCM Strategists invited a number of scholars from around the country to write papers that would discuss the methodological issues in accounting for differences in student populations when evaluating institutional performance. In some cases, these authors were also asked to demonstrate the effects of such adjustments using actual data. In July 2011, these invited authors attended a planning meeting where they and other scholars discussed the planned papers. The planners also invited comments from practitioners and policy experts at that meeting, and another meeting convened in December 2011, when the finished papers were discussed. A major purpose of the second meeting was to shine the light of pragmatism on the work done by the invited authors, in hopes of increasing the chance that the wisdom brought to bear in the studies would have a practical impact in the rough and tumble of policy formation.

## II. USING DATA TO IMPROVE HIGHER EDUCATION

To appreciate the importance of the Context for Success project, it is helpful to recall its particular historical context, which featured growing interest in using data to evaluate postsecondary education. It was in K-12 education, of course, where evaluation-by-measurement came to prominence. Various state policies that evaluated, rewarded and sanctioned schools and teachers using standardized achievement tests found federal legislative form in No Child Left Behind, the landmark Bush Administration law of 2002. It seemed a natural step for this same instinct to find its way into postsecondary education. But it was by no means clear how data would actually be used for the comparatively complex environment of postsecondary training.

### The Push for Measurement and Accountability

The early 2000s witnessed a growing chorus decrying the shortcomings of American higher education. Although there was no single, galvanizing event comparable to Sputnik in 1957, evidence of rising costs, high dropout rates and stagnating rates of college completion fueled concerns. Once a world leader in educational attainment, the U.S. was falling behind in international rankings, as illustrated by a 2007 comparison of attainment among the young and old across countries. Whereas the United States' 55- to 64-year-olds had the highest rate of attainment of postsecondary degrees among 30 Organization for Economic

Cooperation and Development (OECD) countries, its 25- to 34-year-olds ranked only 10th (OECD, 2009). Because of the link between education and workforce productivity, such stagnation in educational attainment represents a direct threat to the continued growth in productivity of American workers. A report from the National Academy of Sciences put it this way:

> The United States takes deserved pride in the vitality of its economy, which forms the foundation of our high quality of life, our national security, and our hope that our children and grandchildren will inherit ever-greater opportunities. That vitality is derived in large part from the productivity of well-trained people and the steady stream of scientific and technical innovations they produce. Without high-quality, knowledge-intensive jobs and the innovative enterprises that lead to discovery and new technology, our economy will suffer and our people will face a lower standard of living. (U.S. National Academy of Sciences 2007, p.1.)

Some critics blamed colleges and universities, believing they had become too self-satisfied and too slow to adopt reforms that would enhance their efficiency. In 2006, a commission led by Secretary of Education Margaret Spellings issued a report castigating American higher education for not taking advantage of technological advances to increase its productivity. To remedy the situation, the commission called for "a robust culture of accountability and transparency." The commission called on institutions to develop "new performance benchmarks designed to measure and improve productivity and efficiency" (U.S. Department of Education 2006, pp. 14, 19, 20). This report was not the first to sound the alarm about the need to improve postsecondary education, nor was it the first to advocate disclosing and using data on outcomes, but its clarion call for quantitative "outcomes-focused accountability systems" (p. 23) famously captures the instinct for measurement that motivated the Context for Success project. Likewise, another national blue ribbon commission of the same period, the National Commission on Accountability in Higher Education, issued a report in 2005 that made recommendations very much in the spirit of the Spellings Commission:

> The National Commission believes we must regain the initiative by holding ourselves to the highest standards of accountability for student success, research and service, and greater productivity in higher education….
>
> We need a fresh approach to accountability, an approach that yields better results. We need accountability to focus attention on state and national priorities and challenge both policymakers and educators to shoulder their share of the responsibility for achieving them. We need accountability to give us dependable, valid information to monitor results, target problems, and mobilize the will, resources, and creativity to improve performance.[3]

3 National Commission on Accountability in Higher Education (2005, p. 6).

These themes have figured prominently as well in debates involving states, accreditation agencies, foundations, advocacy groups and think tanks. Policy analysts and others who have engaged in these debates have grappled with many theoretical and practical challenges in implementing policies that would answer the Spellings Commission's challenge. And in the background of this debate has been the nation's experience with accountability policies applied to K-12 schools, symbolized best by the federal No Child Left Behind. Most observers agree that, whatever its challenges at the K-12 level, the accountability approach will be doubly difficult when applied to postsecondary institutions.

### The Rationale for Quantitative Evaluation

But how, exactly, are transparency, measurement and accountability supposed to improve postsecondary education anyway? Before we can judge the usefulness of adopting quantitative measures of educational outcomes, let alone correcting for inputs, we need to understand what good it will do to have these measures at all. Underlying most reform proposals lie theories of how the world works, with implications about how the proposed reforms will bring about change for the better. In the case of postsecondary education, there exist narratives, usually unstated, explaining why colleges fall short in providing the level or quality of education that could feasibly be produced. These narratives—which are, by the way, similar to those applied in K-12 education, health care and other human service areas—share at least one common theme, called the principal-agent problem.

The principal-agent problem applies to situations in which the person who wants something must entrust to another agent the responsibilities of getting it. According to the textbook economic model, competitive markets can usually be relied upon to ensure that inefficient producers disappear, since the poor quality or high price of their outputs will cause consumers to turn away from them, forcing them to shape up or go bust. But this market mechanism fails to work in markets where the principal-agent relationship is dominant. The problem is asymmetric information. Some services often have more than one dimension, and their quality is hard to measure. To complicate things further, such services are typically produced by complex organizations. Colleges and their workers know what goes on inside classrooms and offices, but the consumers do not. The result is a market where consumers know, as in the aphorism of Oscar Wilde, the price of everything but not the value. Consumers in these cases must depend upon those who work in the organizations providing the services—such as teachers and administrators—to carry out their duties to provide the desired outputs. But slippage may occur, especially if the aims of the employees (the agents) are not identical to those of the ultimate consumers (the principals). One form of slippage might simply be shirking behavior, wherein workers goof off when not being observed. Perhaps more common is the possibility that some agents—for example, professors in the case of colleges—devote more attention to their own research interests, at the expense of the teaching that is the aim valued by students and their parents.

For many critics of postsecondary institutions, this principal-agent narrative both explains the problem and suggests a solution. It explains why colleges might be failing to operate efficiently, because administrators are taking their eye off the ball, which is to teach students. The solution it suggests is to provide information to consumers, so they can make more intelligent choices, and to funding agencies, so they can know where best to allocate their dollars. That consumers of postsecondary education *want* more information was proven beyond a doubt by the instant success of the *U.S. News and World Report* rankings of colleges, introduced in 1983.[4] Although its methodology has changed slightly over time, the rankings it published have been based on a quantitative amalgamation of data on expenditures, admissions selectivity, standardized test scores, retention and reputational ratings. These rankings are universally disdained by the educational establishment—as simplistic and unreflective of important immeasurables—but embraced by the consuming public (and by colleges that brag about being highly ranked). Among the criticisms leveled by scholars is that the *U.S. News* methodology rewards high-priced, wealthy institutions with well-prepared applicants, not necessarily those that produce the most "value added" in their students. Thus the interest in judging postsecondary institutions on the basis of outcomes rather than expenditures and in correcting those outcome measures for the quality of student "inputs."

Quite apart from worries about the technical efficiency of institutions, there has also been ongoing concern about economic inequality in college enrollment and completion. Although rates of college enrollment have been rising over time, rates for those from low-income families have remained lower than those for the affluent. The share of Americans born between 1979 and 1982 who completed college differed markedly by income level. Among those from families in the top income quartile, 54 percent finished college by age 25. By contrast, the rate of college completion was just 32 percent in the third quartile, 21 percent in the second and only 9 percent in the lowest income quartile (Bailey and Dynarski 2011, Figure 4).[5] Even among students with similar levels of academic readiness, those from affluent families are most likely to enroll in college.[6] These disparities in college completion reflect the fact that low-income students are more likely to drop out or never enroll in college, compared with otherwise similar students with more resources. Community colleges, a common destination for many high school graduates of limited means, where many students begin their studies with remedial courses, feature notoriously high dropout rates and low rates of transfer to four-year institutions. Because of the close connection between educational attainment and lifetime earnings, such disparities in college completion portend limited economic opportunities for those already at the bottom of the economic ladder and continued economic stratification in the country as a whole.[7]

4 U.S. News and World Report, "U.S. News History," http://www.usnews.com/usnews/usinfo/history.htm, 1/9/12.
5 These disparities are not new. Among those born between 1961 and 1964, the comparable quartile college completion rates were 61 percent, 46 percent, 43 percent and 26 percent (Bailey and Dynarski 2011, Figure 4).
6 Kane (2004, Table 8.2, p. 333) shows rates of enrollment in postsecondary education and four-year institutions as a function of parental income and math test score.
7 For a study of the connection between education and earnings, see Hanushek.

Contributing to these longstanding gaps in educational attainment, budget cuts in the past decade have been especially severe at public institutions that serve many students of modest means.

### The Logic of Input Adjustment

Concluding that postsecondary institutions should be evaluated by using quantitative outcome measures is one thing. Deciding what measure to use is quite another. Three kinds of measures have received the bulk of the attention: graduation rates, labor market outcomes and test-based assessments. The graduation rate, at its simplest, is the percentage of students who eventually get the degree they aimed to obtain at the outset. One general objection to graduation rates as an outcome measure is that they do not account for quality differences across institutions. Thus a degree from MIT is given no more weight than one from any other four-year college. The second candidate for measuring postsecondary education outcomes is the success of graduates in the labor market, usually reflected by average earnings. To the extent that the labor market rewards better-educated and therefore more productive graduates with higher rates of earnings, this measure may be superior to graduation rates, although it requires more data. The third class of candidate measures is based on direct assessment, via tests, of skills and knowledge acquired by students. I return to the subject of alternative outcome measures below.

There is no better illustration of a quantitative outcome measure—both the motivation for developing it and the logic of correcting it for inputs—than the so-called Student Right to Know (SRK) graduation rate. When Congress was debating a law that would mandate its use, in 1990, Senator Edward Kennedy, a supporter of the bill, echoed the narrative discussed above justifying measurement. He argued that transparency would drive improvement: "Sunlight is the best disinfectant. Once colleges begin disclosing this vital information, those with the poorest records will be under the greatest pressure to improve."[8] Indeed, the law provided for comparative information about postsecondary institutions to be made public, and today its fruit is displayed on the website of the National Center for Education Statistics. A wide array of comparative statistical information is provided about each institution, including enrollment, full- and part-time faculty, graduate assistants, tuition and fees, financial aid, net price by income level, gender and ethnicity breakdown of students, majors available, varsity athletic teams, accreditation, campus crime and loan default rates. [9]

Passed by Congress in 1990, the Student Right to Know Act required that institutions calculate and disclose a precisely and uniformly defined graduation rate. This rate was defined as the percentage of first-time, full-time students who graduate within 150 percent of the "normal" completion time for a degree at the institution where they first enrolled. In the case of four-year institutions, for example, this is the percentage of students who graduate within six years. But there were doubts from the beginning about the usefulness of using this rate to compare the effectiveness of institutions. As the bill was debated, one analyst representing

8 Irvin Molotsky, "Congress Pressing Colleges to Give Figures on Crimes," New York Times, October 8, 1990.
9 NCES College Navigator, http://nces.ed.gov/collegenavigator/, 1/9/12.

an association of public universities, noting its unfairness for institutions with mobile or economically disadvantaged students, asked, "What is the meaning of this number?"[10]

The analyst's misgivings can readily be appreciated by considering a few actual graduation rates. Consider, for example, the six-year graduation rates for 10 four-year institutions in Ohio, shown in Table 1. The table shows a wide range of SRK graduation rates, from Oberlin College (88 percent) and Kenyon College (86 percent) at the top to Wright State University (42 percent) and Wilberforce University (27 percent) at the bottom of this group. Significantly, the table also shows that these graduation rates tend to vary rather systematically with the economic status of the students attending these colleges. The highest graduation rates tend to arise where students come from more affluent families, as indicated by relatively small shares of students receiving Pell Grants. By contrast, the lowest graduation rates occur in the universities with the highest shares of Pell recipients, shown in the table. If one knew nothing more about the 10 institutions than the share of their students who qualified for Pell Grants, one would surely have the suspicion that the 10 were not working with comparable student bodies. Further evidence would surely show, of course, that the two liberal arts colleges with the highest graduation rates also had the best-prepared students, as indicated by average SAT scores or high school grades. Taking the graduation rates at face value would imply that Oberlin and Kenyon are the most effective institutions, but such a comparison would ignore obvious differences in their student bodies.

Comparisons that include for-profit institutions and community colleges suggest even more differences across institutions. Graduation rates for bachelor's students at for-profit institutions generally tend to be low in comparison with most of those displayed in Table 1. For example, DeVry University-Ohio in Columbus reported a six-year graduation rate of 30 percent, while the rate for the University of Phoenix, Cleveland Campus was just 8 percent. And SRK graduation rates for two-year institutions also tend to be low. For example, Davis College, a for-profit institution in Toledo, had an overall rate of 16 percent, Owens Community College in Perrysburg reported 10 percent, Sinclair Community College in Dayton 8 percent and Clark State Community College in Springfield 7 percent.[11]

To be sure, the inadequacy of raw graduation rates such as these has not gone unnoticed. At the time the Student Right to Know Act was enacted, Alexander Astin in the *Chronicle of Higher Education* raised the caution flag about using graduation rates to assess institutional quality. Noting the high correlation between graduation rates and such student characteristics as test scores, high school grades and degree aspiration, Astin asserted that "any 'outcome' measure such as an institution's retention rate cannot be understood unless it is judged in the light of that institution's 'input' of students" (Astin 1990). This criticism, and an attempt to correct for it, is embodied in a contemporary report published by Complete College America, an

10 Ibid.
11 http://nces.ed.gov/collegenavigator/?s=OH, 12/27/11.

organization devoted to improving U.S. graduation rates. The report is generously laced with criticisms of the SRK graduation rates and warnings about the need to take other factors into account when comparing the performance of postsecondary institutions. To improve upon the SRK rates, the report uses information collected from states to calculate a host of particular graduation rates, aggregated at the state level, not found on the federal government's website. Graduation rates not only are shown for four-year, two-year and one-year programs but also are calculated, where data are available, for part-time students, transfer students and full-time students by race and for those who received Pell Grants. For the state of North Carolina, for example, the six-year graduation rate for bachelor's degree students was 63.5 percent for full-time students, 23.1 percent for part-time students and 51.8 percent for full-time students receiving Pell Grants. By presenting graduation rates for these different groups, this report in effect accounts for differences in some (student) "inputs." The question explored in the Context for Success project is whether and how to extend this kind of input correction in producing quantitative outcome measures.

### Lessons from Quantitative Performance Evaluation

The idea of using quantitative measures to assess an institution's performance is by no means unique to postsecondary education. Indeed, this approach, dubbed "quantitative performance evaluation" or "performance management," has been employed or proposed in a variety of other areas of service provision, including job training, health, social welfare services and, as noted above, K-12 schooling. In these systems, quantitative measures or performance standards are developed to proxy the desired outcomes, and the various organizations delivering these services are then evaluated and rewarded on the basis of these quantitative comparisons. The purpose of devising the measures and rewarding organizations that are successful in delivering high levels of the proxy is to mimic to the extent possible the way the private market rewards firms with profits when they produce and sell units of the required quality. The aim of this mechanism is to align the interests of the operating organization (the agents) with the objectives of citizens, or the government (the principal).

James Heckman, Jeffrey Smith and their colleagues examined evaluation systems of this type related to job training. They found that service providers subject to quantitative performance evaluation were indeed responsive to the incentives created by the evaluation scheme, demonstrating the potential of such schemes to affect behavior. Unfortunately, not all responses to program incentives are desirable. For example, one possible perverse response is for organizations to "cream skim" to select only the most promising clients, those most likely to produce successful outcomes. More generally, program designers must worry about "gaming" of the system—that is, producing good outcomes according to the quantitative proxy through practices that are in fact unproductive. Similar tendencies have been documented in the wake of NCLB and similar state programs, leading critics to complain about narrowing of the curriculum to those things covered in tests, focusing on test-taking skills, illegitimate ways of excluding low-scoring students and even manipulating school lunch menus.

Three main lessons arise out of these experiences. First, since "what gets measured gets done," avoid incentivizing only some desirable behaviors and not others. Second, avoid rewarding things over which managers have no control. Third, reward changes in rather than levels of performance. Doing the latter will tend to lead to cream skimming.

## III. SUMMARY OF CONTEXT FOR SUCCESS PAPERS

The project solicited seven research papers touching on issues related to input-corrected quantitative measures of postsecondary outcomes. The authors were asked to address various aspects of the input-adjustment issue as it relates to actual quantitative measures of outcomes. Papers were solicited that would apply or discuss each of three major approaches to outcome measurement: progression/completion, labor market outcomes and direct assessment of learning. In addition, the authors of one paper were asked to examine how costs could be incorporated into input-corrected outcome measures.

The outcome measure with the fewest data requirements, and the one that has been used most often in published reports, is graduation rates. It is, of course, the measure enshrined in federal law, thanks to the Student Right to Know Law. It is also the measure that receives the most attention in the papers prepared for the Context for Success project. The first four papers summarized below concentrate mainly on this measure. The fifth paper, by Cunha and Miller, refers to graduation rates, but its principal focus is on labor market outcomes as an outcome measure. The sixth paper, by Porter, takes up the third alternative outcome measure, direct assessments of learning. One cross-cutting issue, cost, is dealt with explicitly in the paper by Kelchen and Harris. They discuss the issue and illustrate ways costs can be incorporated into outcome measures. The last paper deals with the general issue of diversity across institutions.

### Kelchen-Harris

In their paper, "Can 'Value-Added' Methods Improve the Measurement of College Performance? Empirical Analyses and Policy Implications," Robert Kelchen and Douglas Harris use institution-level data for 1,279 four-year colleges and universities. The outcome measure that they use to proxy for institutional effectiveness is the graduation rate—specifically, the six-year graduation rate in 2008–09—but their methodology could be applied as well to other institution-level measures. The authors demonstrate how this outcome measure can be adjusted to account for differences in inputs. They begin by running regressions explaining the graduation rate, where the explanatory variables include measures of students' academic background (including average SAT and ACT scores), students' financial information (including the percentage of students receiving Pell Grants), students' demographic information (including racial composition) and institutional characteristics (including enrollment and Carnegie classification). Using the estimated coefficients in that regression, they calculate for each college a "predicted" graduation rate,

interpreted as the graduation rate that the average college would have produced if it had enrolled the student this college actually enrolled. The authors then define each college's "value added" as the difference between the *actual* graduation rate and this predicted one. In fact, what they calculate is more accurately thought of as an unexplained difference in graduation rates, but in either case the more positive the value, the better. The most effective colleges are those that graduate a higher percentage of students than they are predicted to, based on the characteristics of their students.

To avoid having roughly half of all institutions end up with a calculated "value added" that is negative, surely an awkward thing to explain, the authors recalibrate the calculated values by adding enough to all of them so that the worst-performing college is assigned a value added of zero. Effectively, their measure of value added is interpreted as the contribution made by any college to improving the chance of graduation as compared with the worst-performing college. Making a point that is taken up in other papers as well, the authors emphasize that a measure such as this is superior to most of the measures that go into the *U.S. News* rankings, which are measures of resources expended, academic preparation of students, selectivity or reputation. Their value-added measure reflects production, or contribution to learning, rather than the cost of production or the value of the inputs. This, they argue, is what the public and students and their families care about, or should care about.

They also care about how much it costs to produce these outputs. Therefore, Kelchen and Harris also provide cost-adjusted measures of output. For the public, they propose value added divided by the cost of instruction. Institutions that are able to increase graduation rates by one percentage point at the lowest cost are most efficient. Students and their families will have a different calculus, however, since they end up paying only the net price, tuition and fees minus scholarship aid. The proper ranking for them, the authors argue, is value added divided by net price. Both of these cost adjustments can make a big difference in ranking which colleges are most effective, as the authors illustrate by ranking actual colleges according to their three value-added measures: uncorrected for cost, corrected for total cost and corrected for net cost to families.

### Wright, Murray, Thrall, Fox and Carruthers

In "College Participation, Persistence, Graduation, and Labor Market Outcomes: An Input-Adjusted Framework for Assessing the Effectiveness of Tennessee's Higher Education Institutions," David Wright et al. offer a second paper that demonstrates how regression analysis can be used to produce input-adjusted measures of institutional effectiveness. They employ data for Tennessee, a state that actually uses measures of institutional effectiveness as a significant part of its funding formula for public colleges and universities. Employing data on individual students who matriculated in two- and four-year public institutions in 2002, these authors estimate regressions for two sets of outcome measures. The first set encompasses several measures of progress, such as the SRK graduation rate. The second includes several measures of labor market success, such as the employment rate (in Tennessee) and the wage rate for those employed in the

state. The authors introduce into their estimated equations three sets of explanatory variables, beginning with a set of demographic and educational variables that are commonly found in administrative data sets, such as gender, race, age, test scores and income level. They collected a second set of student characteristics from matching each student's home address to a census tract and using those neighborhood variables. A third set of variables was derived from matching each student's location to one of a dozen broad psychometric marketing classifications, taken from a proprietary marketing data set.

Consistent with the previous study, Wright et al. find that controlling for student inputs in these regression models makes a big difference in the rankings. Among the estimated effects from adding these variables, they find lower graduation rates, for example, for males, those with low test scores, those with low income and those who lived nearby. To illustrate the difference made by the regression-based correction, the college that showed the strongest average graduation rate (.29 above the omitted institution) dropped precipitously after controlling for the first set of student variables, to .09. By contrast, another college, which by comparison was remote and less advantaged, showed an increase, from an uncorrected rate of .17 above the reference institution to an adjusted rate of .28 above it. The college with the relatively affluent student body, which looked so good based on unadjusted graduation rates, was taken down a few pegs when student characteristics of the two colleges were controlled for.

Adding the second and third sets of explanatory variables added little to the explanatory power of student characteristics, however, and therefore had little influence on rankings. For states or other evaluation groups wishing to undertake a correction for inputs, this finding is reassuring, because it implies that readily available student information will do about as good a job as more extensive (and expensive) sets of data. In particular, there is little justification for buying the proprietary marketing product featured in this paper. The key result is that differences among institutions, which are very large before any correction, shrink markedly when inputs are controlled for. After all possible controls are entered, however, significant differences persist, suggesting that some colleges are more effective, or efficient, than others.

### Pryor and Hurtado

In their paper, "Using CIRP Student-Level Data to Study Input-Adjusted Degree Attainment," John Pryor and Sylvia Hurtado offer a third application of regression analysis, this time using data on individuals for whom they have collected extensive information through the CIRP Freshman Survey. Introduced in 1966, the Freshman Survey collects self-reported information from entering first-year students, for colleges to use in comparing their students with those enrolling in similar institutions. To this survey information, the authors have added information on transfer and graduation, using the National Student Clearinghouse. In their examination of graduation rates, therefore, the authors are able to overcome a serious shortcoming of the SRK graduation rate, namely its lack of attention to students who transfer and eventually graduate from a different college.

The authors bring to their analysis one more advantage that will not be available to many researchers—an extensive array of student-level answers to survey questions, measuring personal characteristics, past behavior, current attitudes and expectations for the future. From the 250 student-level variables available to them, they restrict themselves to 76, augmenting these with five institution-level measures that they created based on their survey results ("peer norms") and 11 other institution-level measures that are commonly available.

Their regression results explaining graduation rates have two noteworthy features. The first is the very large number of explanatory variables they use. Their collection of 76 variables, taken from the Freshman Survey, document many facets of high school experiences, attitudes and expectations that are, obviously, not available to researchers who do not have access to these survey results. The authors document the $R^2$ for a regression explaining graduation with this extensive set of variables (.245), the fact that statistical fit is improved somewhat when their five peer norms variables are added ($R^2$ = .262). The clear implication is that adequate correction for inputs requires a large number of measures, such as those available in the Freshman Survey.

The second feature demonstrated by the Pryor-Hurtado analysis is the desirability of having data on transfers. They find that transfers who go on to gain a degree at another institution, their so-called mobile completers, are not a random sample of all matriculants. Having information on them produces a fuller and more convincing outcome measure.

### Bailey

Although Thomas Bailey's contribution, "Developing Input-Adjusted Metrics of Community College Performance," is ostensibly about community colleges, it contains insights that apply more generally to postsecondary education. He begins by emphasizing a fact that bears repeating: that colleges and universities are multifaceted to an extent far beyond what is found in K-12 education. These institutions are a classic case of multiproduct firms, ranging from highly specialized technical and mechanical training to broad and abstract learning. Within postsecondary education, it is the two-year community colleges where most of this variety in aims can be observed. Two other features that complicate analysis of postsecondary outcomes, as opposed to those at the K-12 level, are part-time enrollment and transfers between institutions. Part-time students make up a large share of all college students, especially in community colleges, and students can transfer (without first obtaining a degree) at some point in their educational careers. A further complication is the large number of students in community colleges who enroll in noncredit courses that are not part of any formal degree program. As one very specific example of how these features complicate analysis of outcomes, they all play havoc with any attempt to measure institutional performance by measuring students' graduation, such as with the Student Right to Know graduation rate. Because it is based on a graduation rate only for first-time, full-time, degree-seeking students, Bailey emphasizes, it fails to

account for transfer students, part-timers or those who are not in degree programs. In addition, since the SRK methodology mandates the inclusion of certificates (courses of study involving less course work than associate degrees), measured against proportionately shorter time periods, the methodology implicitly equates degrees and certificates.

Bailey offers a number of ideas for improving the SRK methodology. Most important, he calls for following a broader cohort, one that would include students who transfer to different colleges. Such a change would require what is called "student unit record data" that would go beyond what any one institution could easily collect. The obvious candidate for record keeper for such an expanded graduation rate would be the federal government. Other improvements include better differentiating among certificates of differing lengths and associate degrees and between academic and vocational programs; accounting for differences across colleges in the mix of programs offered; and measuring progression at thresholds below actual graduation. He also makes the case for paying more attention to student intentions, measures of program quality and differences in student characteristics and program costs. As noted below, he advises analysts to "control sparingly" for institutional characteristics as a part of input adjustment and to restrict comparisons to similar institutions. Lastly, he makes a case for emphasizing changes in performance measures for a given institution over comparisons across institutions.

### Cunha-Miller

In their paper, "Measuring Value-Added in Higher Education," Jesse Cunha and Trey Miller provide "a practical guide for policymakers," combining the demonstration of statistical correction for inputs and guidance for using such techniques in a policy framework. Their empirical work employs data for four-year public institutions in Texas. Like Kelchen-Harris, they employ regression analysis with right-hand-side variables measuring student characteristics. Unlike the former paper, their paper employs data for individual students (specifically, data from Texas' longitudinal student information system) rather than institutions. In addition, they focus on an alternative outcome measure, labor market earnings six years after graduation, in addition to graduation rates. For right-hand-side variables, they adopt a modified "kitchen sink" approach, which deliberately includes as many variables as practicable in order to improve on statistical fit.

Instead of comparing outcomes with a predicted standard to arrive at "value added," this paper compares input-corrected outcome levels (of earnings, persistence or graduation) with those achieved by students from a reference institution—specifically, Texas A&M. The input-corrected outcome level, averaged by institution, is the institution fixed effect estimated in the regression equation. Including measures of student background serves the purpose of controlling for input differences, as in the previous paper. The authors point out that these differences arise in large part because students have considerable latitude in choosing where to attend. Whatever the cause, such differences in student characteristics need to be controlled for, lest institutions receive unwarranted credit for imbuing in their students qualities that they possessed before

matriculation. The authors go on to warn of the dangers of bias arising from student characteristics that remain unobserved even in regression models. The authors take steps to account for such bias, called selection bias, in one of their models. Adopting a technique developed in previous research, they present estimates that are based on differences in institutional outcomes based entirely on observing students who had the option of attending more than one of a group of colleges, a subset of all students.

Cunha and Miller's results illustrate the general finding, discussed below, that controlling for inputs generally narrows observed differences, and by a great deal when every last tool is used to ensure an apples-to-apples comparison, as shown in Table 3. These results strongly suggest that unmeasured student characteristics differ systematically across colleges, pushing up the conventionally measured superiority of colleges attended by well-prepared students. When measured inputs are controlled for, differences between colleges shrink, and when the authors seek to control for selection effects, the differences become very small indeed. The table also shows that these adjustments may change the ranking of institutions, but many of the differences between institutions are too small to be statistically reliable, especially in the models where the differences shrink and become almost imperceptible.

### Porter

As the title suggests, Stephen Porter's paper, "Using Student Learning as a Measure of Quality in Higher Education," concentrates on the third general type of assessment measure— direct assessments of student learning. Rather than infer from graduation or labor market earnings what a student learns in college, it is possible to assess learning directly, from a test or survey taken by students. Prominent examples of tests used to assess postsecondary learning are CAAP (Collegiate Assessment of Academic Proficiency), MAPP (Measure of Academic Proficiency and Progress, or Proficiency Profile) and CLA (Collegiate Learning Assessment). Alternatively, students can be asked about their learning or about features of their student experience that are thought to be correlated to learning, an approach taken in NSSE (National Survey of Student Engagement). The test of a good test, writes Porter, is how closely correlated it is to actual student learning, how representative institution-level averages are of students, and how well those averages account for differences in student ability across institutions. The last of these three criteria, of course, is the goal that is meant to be served by input adjustment.

Porter offers his views on the advantages and shortcomings of the above-named assessment instruments. For the NSSE, for example, he notes problems with self-reporting, including problems with recall, vaguely worded questions and psychological biases. For tests of knowledge, like CAAP and MAPP, there is little chance to synthesize or apply knowledge. By contrast, the CLA focuses on skills of discernment and analysis, but the CLA has low response rates, in large part because it is a lengthy test. Selling points of CAAP are that it has strong content validity, it has a relatively strong correlation with grades and it is comparatively cheap and easy to administer. Porter argues that value added is best measured by comparing outcomes of such

tests at the beginning and end of college, for the same students, but attrition rates can be high, making such comparisons impossible.

### Bahr

In "Classifying Community Colleges Based on Students' Patterns of Usage," Peter Bahr offers a numeric analysis that underlines the "multifaceted mission" of community colleges. As his title suggests, Bahr classifies community colleges based on the prevalence among each college's students of six mutually exclusive enrollment patterns that he identified in previous research. These patterns drive home the reality that community college students enroll for different reasons and have different patterns of course-taking during their period of enrollment. Some students stay on track and receive degrees while others seem to jump in and jump out. Some focus on obtaining technical training while others appear intent on transferring to a four-year institution. Using data from 105 community colleges in California, Bahr differentiates colleges based on the prevalence of his six types of students. He then compares these institutions by such classifications as location, funding and financial aid. Regardless of the precise details of his analysis, the paper represents an emphatic caution against "one size fits all" thinking when it comes time to design empirical measures of institutional performance.

## IV. THEMES

In summarizing these papers and the discussion that occurred at the two meetings of participants, I find it useful to begin with points that appeared to enjoy such wide acceptance that they did not require stating explicitly.

### Unspoken Points of Agreement

Some of the most important points of agreement are beliefs that are shared so completely that they are never or rarely stated explicitly. In spite of their absence in discussions, or perhaps because of that, these beliefs are worth spelling out explicitly because they constitute an important part of the intellectual foundation for debate. In this category, I note five points of agreement, only some of which were ever mentioned explicitly during any of the discussions that occurred as part of this project.

Resources are scarce, so they must be husbanded. This is a foundational principle of economics, and it also underlies the rationale for evaluation of postsecondary education. The implicit objective for society in postsecondary education, as it is in the textbook example of any product, is to maximize the output of some valued outcome subject to the limited resources that can be devoted to it. In the case of education, this output involves learning certain skills and acquiring certain knowledge. Despite the vagueness of these words, assume this outcome can be measured. Call this Q. This ultimate objective is followed by three allied principles: (1) Allocate resources to the institutions and programs with the biggest payoff in terms of Q per

dollar. At the margin, a state should allocate its budget so that the marginal increase in Q per dollar is the same at every one of its institutions. This is, of course, the standard optimization rule found in any microeconomics textbook. (2) Create institutional incentives that will accomplish this objective. This means, don't reward the wrong things—those pesky unintended consequences that we have seen so much of in K-12 assessment schemes. (3) Design your policy so that the implementation itself will also be cost-effective. As we move from objectives to implementation, this principle is important, and it is especially highlighted in the Cunha-Miller paper. It implies, for example, thinking hard about the data you'll need to carry out your assessment policy, including whether you'll need to have data on each student.

Higher education has multiple, amorphous outputs. Despite the agreement on this waste-not, want-not economist's orientation, there is not general agreement about how to measure, and this is largely caused by the multitude of things that colleges do. To an extent not matched at the K-12 level, postsecondary institutions do many things. Even putting aside research and service, the kinds of teaching done in college vary widely, from the practical and job-related to the theoretical and other-worldly. This diversity in subject matter does not capture the full range of things that are taught and learned in college, however. College students are expected to grow in their ability to discern relevant questions to pursue, analyze problems and make cogent arguments. The complexity of these various objectives is a daunting challenge to measurement of outcomes, summarized by the variable Q above.

In devising proxies for learning, scholars have used three kinds of outcome measures: graduation or progression rates, labor market outcomes and direct tests of learning. Of these, the old standby, graduation rates, is probably the easiest to implement and defend, but it is not without its problems, as several authors and discussants pointed out. The biggest drawback is that degree attainment says not the first thing about content or quality. Using it as an outcome measure also invites unintended effects in the form of watering down requirements and quality in order to push more students out the door with a degree in hand. There are other practical problems with this measure as well, such as penalizing institutions with many part-time students or those whose students transfer to other colleges to complete their degrees. Labor market outcomes have the virtue of letting employers assess the quality of instruction with their paychecks, but these measures have problems of their own. They are available only for those who enter the labor force, they aren't too meaningful in the first few years after graduation, and they completely miss many important dimensions of education that have no economic payoff. Direct measures of learning are at once most attractive and most problematic, requiring great expense and differentiation if done correctly. These three approaches all have their drawbacks, and the participants in this project saw the tradeoffs among them largely in terms of cost and authenticity.

Education of the disadvantaged requires attention. Any close study of American education will reveal that students who grow up in low-income families tend to have lower educational attainment than their more

affluent peers with comparable achievement levels. As noted above, less than a tenth of those who grew up in the nation's bottom income quartile finish college by age 25, compared with more than half for those from the highest income quartile. Clearly, this fact does not square with the ideal of equality of educational opportunity, that a person's educational attainment should not be determined by circumstances over which he or she had no control. Beyond equity, however, attention to the educational progress of low-income students can be justified by the same precepts of efficiency noted under the first point of this section. The evidence would suggest, in short, that too few of the nation's resources have been devoted to the education of those who grew up attending schools in low-income communities and neighborhoods.

Second best is better than third best. In designing measures and statistical procedures for use in accountability schemes, the authors of these papers reveal a healthy regard for practical limitations. These papers admirably illustrate the admonition implicit in Voltaire's famous phrase "The best is the enemy of the good."[12] The authors of these papers are not hung up on specifying the best way to measure the outcomes of postsecondary institutions and the many ways their measures fall short of the best. Rather, they are going about the pragmatic business of designing measures and procedures that might actually be implementable. The world they perceive is an imperfect one, where governments don't know enough about people or the process of education to specify all these models perfectly, but where events do seem to be heading in the direction of some type of measure being used, whether or not policy analysts are fully prepared.

Beware of unintended consequences. Although it is really part of the first principle in this list, this principle deserves repeating. Unlike the other principles, this one was explicitly stated, by several authors and participants, as an admonition to those who would design policies. It is embodied in the axiom "What gets measured gets done." It is best, therefore, to design policy so the incentives of actors are in line with the desired outcome. As analysts of K-12 schools are only too aware, the path of No Child Left Behind and similar state assessment programs is littered with unintended consequences, ranging from narrowing the curriculum to devious steps taken to keep low-achieving students from being included in assessments. At the postsecondary level, one feared unintended consequence of heightened emphasis on graduation rates, for example, would be a fall in academic standards for graduation.

### Consensus Lessons

Other points of agreement were less axiomatic, but rather emerged as consensus judgments that grew out of the papers and discussions. They are most conveniently stated in the form of lessons, warnings to those who would venture into the terrain of quantitative measurement and correcting for inputs.

12 Le mieux est l'ennemi du bien, which can also be translated as, "The better is the enemy of the good." Voltaire, "Dramatic Art," Dictionnaire philosophique (1764).

Distrust average outcomes for institutions. If there is one point of agreement that arose from this project, consistent and clear, it is that average outcomes at the institution level, no matter what measure is used, are a misleading guide in identifying which colleges are the most effective. This conclusion corresponds exactly to the track analogy. If differences in students' starting points are not taken into account, there can be no assurance that average outcome proxies reflect differences in institutional effectiveness rather than differences in the academic readiness of entering students. It also is fully consistent with the logic of equating effectiveness with the concept of value added. Controlling for differences in inputs is important, and any comparison that does not do this should not be taken at face value.

Account only for predetermined inputs. The reason to control for student characteristics is to take account of differences in students' starting points. One might well expect that outcomes will be affected by practices undertaken by colleges, such as extra counseling or small classes, or by the quality of faculty or facilities. Controlling for practices or features such as these is not appropriate because these are the very stuff of institutional effectiveness. For the same reason, it would make no sense to control for differences, for example, in the amount of practice time each runner undertakes in training for the race.[13]

Follow students who transfer. From a state or national perspective, there is little reason to applaud a student who graduates from the college where he or she first enrolled any more than a student who transfers and then graduates from the second college. Assigning credit to each of the colleges in the latter case might present difficulties, but it would be foolish to use a proxy for a successful outcome that gives credit to the former student but not the latter. Unfortunately, that blind spot is one of the features of the SRK graduation rates, since it makes its calculations only for students who remain in the same institution through to degree attainment. The obvious remedy is to keep track of students when they transfer between colleges, a goal that would be achieved if data were collected in so-called student unit records.

Make comparisons among similar institutions. One rough-and-ready way to deal with the analytic challenge arising from the wide variation across colleges is to restrict the set of colleges being compared. Doing so deals with—albeit incompletely—two kinds of differences: in the readiness of students to learn (the controlling-for-inputs problem) and in institutional missions (the multidimensionality problem). Intuitively, this approach makes sense, and it has the virtue of not requiring statistical operations that will have the look and feel of a black box. In comparing the Ohio colleges shown in Table 1, for example, this approach says that, if one is using graduation rates as an outcome measure, it is more sensible to compare Oberlin and Case-Western than Oberlin and Kent State.

[13] Although several authors made this recommendation, Thomas Bailey stated it early and emphatically in his papers, advocating with understatement to "control sparingly" for measures describing colleges. Thus this rule came to be referred to as the "Bailey axiom."

To be clear, this approach is not an absolute rule, but rather a pragmatic response to the desirability of controlling for differences in inputs and mission. If one believes that an entire category of institution is in need of change or improvement, restricting comparisons to this group would not be ideal.

Account for costs. A sine qua non of efficiency is attention to the cost of operation. If College A produces the same value-added learning as College B, but at 80 percent of the cost, College A is more effectively using its resources. None of the authors would advocate ignoring costs. But there was less agreement about how costs ought to be accounted for. Kelchen and Harris, for example, recommend dividing their measure of value added by the cost per student (full costs or private costs, depending on whose perspective is taken). If one questions their measure of value added, however, such a simple calculation would also be subject to question. It may be wiser merely to present data on costs alongside estimates of value added and allow the consumers of the information to make the comparison themselves.

Proceed on several fronts. In light of the imperfections attached to all of the measures and approaches under consideration, it would seem wise not to be wedded exclusively to one path. The advantage of using persistence and graduation rates is that they are comparatively easy to collect and are in fact already available, thanks to the Student Right to Know Law. But, as discussed above, there is much to be said for looking at labor market outcomes, since these arguably reflect quality. Another way to get at quality differences across institutions is through directly testing students. This is likely to be the most expensive of the three approaches. Depending on which of these measures is feasible and the cost of collecting and analyzing the results, there would be much to gain from taking an eclectic approach. If alternative approaches point in the same direction, analysts can take some comfort. If not, the reverse will be true.

### Two Big Questions

Near the conclusion of the second meeting of authors and commentators, the assembled group took the time to debate two policy questions that had been hanging over the proceedings, unresolved.

Is it time to incorporate input adjustments into performance evaluation of colleges?

As noted by one participant, this question has two parts, one having to do with performance evaluation and one having to do with input adjustment. As to the first part, participants stressed the value of doing evaluations using measures that institutions themselves do not control. This is a central component of No Child Left Behind, and one good result of that law has been to illuminate the significant gaps that exist between groups of students. For all its flaws, they argued, the very act of assessment using a comparable yardstick has had real benefits. As to the input correction, the strongest argument on the affirmative side is the consensus belief, noted above, that outcome measures *should* be corrected for differences in student inputs and that simple averages calculated at the college level are going to be misleading.

But will the cure be worse than the disease? Because the available outcome measures—in particular, graduation rates—are so crude, they are weak on two counts. Their imperfections will undercut support for them and, by extension, the evaluation process itself. And they will be vulnerable to gaming by colleges in ways parallel to responses that have been observed in other service delivery areas. To illustrate the potential for this, one participant summarized a view taken by some observers of community colleges that policies using graduation rates to allocate resources would quickly result in a lowering of standards for graduation.

In balancing these pros and cons, the group assembled for the project's second meeting was not unanimous, but I judge that the nays outnumbered those who think the policy community is ready to implement evaluation using input-adjusted outcome measures. The majority felt that input-adjusted measures now available are too flawed to be incorporated in full-blown data-based performance evaluation. That is not to say that no evaluations should be made using quantitative metrics. Indeed, they are already being made. Where that is happening, it makes sense to do what can be done to use sensible measures, sensibly adjusted. A pragmatic approach, then, is to adopt a limited, modest approach to comparing institutional effectiveness, one with fewer data demands. This approach would limit comparisons only to those among similar institutions. As noted above, this kind of restriction implicitly serves as its own rough-and-ready correction for inputs, to the extent that similar institutions enroll similar students. Although no precise groupings are likely to be agreed on, comparisons among reasonably similar colleges should be more meaningful than, say, comparing the graduation rates of Oberlin and Kent State. Even if such comparisons are not employed explicitly in formal performance evaluation systems, data could be published for the use of potential consumers, in the spirit of *Consumer Reports* product comparisons. The NCES College Navigator, which currently provides information that parents and students can access, would be the type of platform on which such comparative data might be presented.[14]

What should be the top priorities for improving data? As with the first question, the discussion of this second question implicitly recognized that policymakers must weigh the good and bad points of various approaches. And again the assembled group was not unanimous in its views. But two points did seem to enjoy widespread support. First, to the extent possible, we should take advantage of data sets that are already developed. Second, because of the prevalence of student transfers between colleges, a student record system that follows students is the way to go. Some thought this should be a federal system, while others thought that a great deal of improvement would be achieved simply by improving state record systems.

14 Data are also made available by the state of Texas through the Texas Higher Education Accountability System, which provides online, accessible information.
http://www.txhighereddata.org/interactive/accountability/, 12/27/11.

Near the end of the discussion at the second meeting, one participant spoke up with a reminder that may well be a fitting admonition with which to conclude. As we consider the advantages and disadvantages of putting together measures of outcomes and evaluation systems based on them, it is important to keep in mind the audiences or potential consumers of these measures. Some of the potential consumers will be quite sophisticated users of data, among them state coordinating boards, system administrators, governing boards, and some citizen and advocacy groups. What is an appropriate level of complexity for these groups may not be so if the aim is to put information out for parents and students to use.

### Aligning Purpose and Method

The Context for Success project was from beginning to completion about methodology—about methods of correcting quantitative measures of educational outcomes for differences in student inputs in order to make those measures more meaningful and useful. But underlying this methodological discussion is the ultimate aims for which the measurement was devised in the first place. This distinction among aims is a point made by several participants in the project, and it is worth reiterating as I conclude this synopsis.

Quantitative outcome measures such as those discussed in this project have at least three distinct kinds of potential consumers, each wanting information for their own particular purposes. These potential data users are students (and their parents), funders and other decision makers, and the colleges themselves. Students and their parents are understandably hungry for information that can help them in what is often the biggest investment they will ever make. Whether or not a student borrows money, the cost of postsecondary education, in the form of bills to be paid and earnings that are forgone, can be frightening. Like consumers contemplating other major purchases, from refrigerators to cars, these consumers desire information and, as indicated by the sales of the annual *U.S. News* college edition and other college guides, they are willing to pay for it. So, in the spirit of *Consumer Reports*, the Education Department's College Navigator makes information available online about graduation rates and other useful facts about colleges. Rankings by *Washington Monthly* are also based, in part, on graduation rates.[15]

A more sophisticated set of consumers are the government agencies, governing boards, funders and accreditation bodies that seek objective ways to rate, rank and reward institutions based on their effectiveness in educating students. Because they are professionals, presumably well versed in the uses and shortcomings of quantitative measures, there should be less reason to worry about the complexity that input adjustment inevitably brings with it. Clarity is to be desired, to be sure, but simplicity is not the imperative it must be for measures designed for students and their families.

[15] The *Washington Monthly* methodology involves comparing actual graduation rates with predicted rates, much in the spirit of the paper by Kelchen and Harris. The predicted graduation rates are based on a regression of graduation rates on SAT scores and percent of students getting Pell Grants. See Washington Monthly, http://www.washingtonmonthly.com/college_guide/rankings_2011/national_university_rank.php 2/19/12.

The third group of users are the colleges themselves. Who should be more interested than they in finding out which programs work and which do not? In seeking to improve what they do, colleges may well want to apply these methods to programs, not the entire institution. But however they decide to use quantitative measures, it seems they ought to want to do it. In his paper, Porter argues persuasively that colleges and universities devote far too few resources to assessing the learning of their students. To be sure, there is already plenty of assessment going on in higher education. But he is referring to assessments over and above this ubiquitous testing and grading regime. He is advocating evaluation that will measure learning in a way that is a closer approximation of the value-added concept, one that is based on a yardstick that can be understood by knowledgeable observers besides just the instructor and the students. Those familiar with colleges and universities will understand that this kind of evaluation is rare. In most stand-alone courses, assessment of learning is the province of the instructor alone. But Porter makes a compelling case that this should be a much more pressing concern of institutions themselves, and that they should put more resources into evaluation other than the conventional, instructor-controlled variety. The professorate instinctively recoils from any suggestion of standardized assessment, for reasons legitimate and otherwise, but more serious attention to this kind of assessment is needed, given the scarcity of resources and the national interest in using those resources effectively. In this spirit, accreditation bodies have been asking institutions to devise new ways of assessment, and there is every reason for faculty and administrators to view this as a welcome challenge rather than unproductive interference.

# TABLES

| Table 1 | | |
|---|---|---|
| Student Right-to-Know Graduation Rates, | | |
| Selected Four-Year Institutions in Ohio, 2011 | | |
| | | |
| Institution | 6-year graduation rate for students pursuing bachelor's degree | Percentage of full-time beginning students getting Pell Grants |
| Oberlin College | 88 | 10 |
| Kenyon College | 86 | 8 |
| Case-Western Reserve University | 82 | 20 |
| Miami University (Oxford) | 80 | 14 |
| Ohio State University | 78 | 18 |
| Ohio University | 65 | 24 |
| University of Cincinnati | 56 | 26 |
| Kent State University | 50 | 33 |
| Wright State University | 42 | 38 |
| Wilberforce University | 27 | 80 |
| | | |
| Source: National Center for Education Statistics, College Navigator, http://nces.ed.gov/collegenavigator/?s=OH, 12/27/11. | | |

Table 2

| | Table 2 | | | |
|---|---|---|---|---|
| | | | | |
| Six-Year Graduation Rates, Selected Master's Institutions | | | | |
| | | | | |
| | Rank based on Kelchen-Harris value-added | SRK 6-year graduation rate (%) | Rank based on SRK rate | Percent Pell (%) |
| The Citadel | 1 | 68 | 4 | 20 |
| College of Notre Dame of Maryland | 2 | 62 | 7 | 25 |
| Penn State - Harrisburg | 3 | 78 | 2 | 29 |
| Rosemont College | 4 | 70 | 3 | 59 |
| Gwynedd-Mercy College | 5 | 65 | 5 | 28 |
| College of St. Elizabeth | 6 | 59 | 8 | 26 |
| Rutgers University - Camden | 7 | 58 | 9 | 37 |
| University of Illinois - Springfield | 8 | 57 | 10 | 32 |
| Regis College | 9 | 64 | 6 | 33 |
| College of New Jersey | 10 | 85 | 1 | 16 |
| | | | | |
| (a) For students entering in fall 2002 | | | | |
| | | | | |
| Source: National Center for Education Statistics, College Navigator, | | | | |
| http://nces.ed.gov/collegenavigator/?s=OH, 12/27/11; | | | | |
| Kelchen and Harris, Table 6. | | | | |

**Table 3**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Table 3 | | | | | | |
| | | | | | | | | | |
| | Input-adjusted Outcomes for Selected Institutions, Log Earnings and Graduation Rates, Difference from Texas A&M, Cunha-Miller | | | | | | | | |
| | | | | | | | | | |
| | ----------Log earnings---------- | | | | | | | Graduation rate | |
| | | Rank on | | | | | Rank on | Rate | Rank |
| | (1) | (1) | (2) | (3) | (4) | (5) | (5) | | |
| U. Texas Tyler | -0.24 | 5 | -0.23 | -0.15 | -0.13 | 0.01 | 1 | -0.18 | 9 |
| Texas Women's College | -0.30 | 9 | -0.21 | -0.12 | -0.08 | 0.00 | 2 | -0.10 | 7 |
| Texas A&M U. | 0.00 | 1 | 0.00 | 0.00 | 0.00 | 0.00 | 3 | 0.00 | 3 |
| Texas Tech U. | -0.18 | 4 | -0.18 | -0.07 | -0.05 | 0.00 | 4 | -0.03 | 4 |
| S.F. Austin State U. | -0.28 | 7 | -0.25 | -0.16 | -0.12 | -0.01 | 5 | -0.10 | 7 |
| Tarleton State U. | -0.28 | 8 | -0.29 | -0.14 | -0.10 | -0.01 | 6 | -0.08 | 6 |
| U. Texas Dallas | -0.12 | 2 | -0.12 | -0.14 | -0.11 | -0.02 | 7 | -0.06 | 5 |
| U. Texas Austin | -0.13 | 3 | -0.12 | -0.12 | -0.11 | -0.06 | 8 | 0.04 | 2 |
| Lamar U. | -0.35 | 10 | -0.32 | -0.21 | -0.16 | -0.08 | 9 | -0.20 | 10 |
| U. of Houston | -0.26 | 6 | -0.21 | -0.20 | -0.17 | -0.11 | 10 | 0.15 | 1 |
| | | | | | | | | | |
| Models based on log earnings of graduates 8 years after h.s. graduation | | | | | | | | | |
| (1) Uncorrected | | | | | | | | | |
| (2) Controls for race and gender | | | | | | | | | |
| (3) Adding high school fixed effects and courses taken in high school | | | | | | | | | |
| (4) Adding SAT score and demographic information available from SAT | | | | | | | | | |
| (5) Adding application group fixed effects | | | | | | | | | |
| | | | | | | | | | |
| Source: Cunha and Miller, Table 2. | | | | | | | | | |

# REFERENCES

Achieving the Dream, *Test Drive: Six States Pilot Better Ways to Measure and Compare Community College Performance* (Jobs for the Future, n.d.) http://www.achievingthedream.org/_pdfs/_publicpolicy/testdriveXS.pdf, 10/21/10.

Astin, Alexander, "Retention-Rate Data Mislead Student 'Consumers,'" *Chronicle of Higher Education*, November 21, 1990.

Bailey, Martha J. and Susan M. Dynarski, "Gains and Gaps: Changing Inequality in U.S. College Entry and Completion," NBER Working Paper 17633.

Barnow, Burt S. and Jeffrey A. Smith, "Performance Management of U.S. Job Training Programs: Lessons from the Job Training Partnership Act," *Public Finance and Management* 4 (2004), 247-287.

Complete College America, *Time is the Enemy* (Complete College America, 2011). http://www.completecollege.org/state_data/, 12/29/11.

Hanushek, Eric A. and Ludger Woessmann, "The Role of Cognitive Skills in Economic Development," *Journal of Economic Literature* (September 2008), 607-668.

Heckman, James J., Carolyn J. Heinrich, and Jeffrey Smith, "Performance Standards and Potential to Improve Government Performance," in James J. Heckman, Carolyn J. Heinrich, Pascal Courty, Gerald Marschke and Jeffrey Smith (eds.), *The Performance of Performance Standards* (Kalamazoo, MI: W.E. Upjohn Institute for Employment Research, 2011).

Kane, Thomas J., "College-going and Inequality," in Kathryn M. Neckerman (ed.), *Social Inequality* (New York: Russell Sage Foundation, 2004), pp. 319-354.

Muriel, Alastair and Jeffrey Smith, "On Educational Performance Measures," *Fiscal Studies* 32 (2011), 187-206.

National Commission on Accountability in Higher Education, *Accountability for Better Results* (State Higher Education Executive Officers, March 2005). http://www.sheeo.org/account/accountability.pdf, 12/27/11

U.S. Department of Education, *A Test of Leadership: Charting the Future of U.S. Higher Education* (Washington: Department of Education, September 2006), http://www2.ed.gov/about/bdscomm/list/hiedfuture/reports/pre-pub-report.pdf, 10/21/10.

U.S. National Academy of Sciences, Committee on Science, Engineering, and Public Policy, *Rising above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future (Washington, D.C.: National Academy Press, 2007).* http://www.nap.edu/openbook.php?record_id=11463&page=1, 12/28/1

*Context for Success is a research and practice improvement project designed to advance the best academic thinking on postsecondary institutional outcome measures. The project was organized by HCM Strategists LLC with support from the Bill & Melinda Gates Foundation. The papers may not represent the opinions of all project participants. Readers are encouraged to consult the project website at: www.hcmstrategists.com/contextforsuccess.*