Buyer beware:
Why consumer-oriented comparisons of colleges and universities mislead and distort

David Radwin

MPR Associates, Inc.

Presented at the annual conference of the California Association for Institutional Research

Sacramento, CA, November 20, 2009

In its 2006 report, the Commission on the Future of Higher Education appointed by then-Education Secretary Margaret Spellings called for a "consumer-friendly database that provides [public] access to institutional performance and aggregate student outcomes" to help prospective college students and their parents make better-informed choices.[1] Colleges and universities have responded with the Voluntary System of Accountability[2], the University and College Accountability Network[3], and similar endeavors[4] to present an array of uniformly formatted statistics that encourage comparisons between institutions. These online databases combine widely available information, such as graduation rates, tuition, and admissions statistics, with previously obscure or non-public measures of academic engagement and learning outcomes. Although these websites invite the public to comparison shop for colleges as they would for products like automobiles or toasters, institutions of higher education are not consumer goods, and pretending otherwise inevitably invites error and misinterpretation. This bias is most clearly evident in comparisons of undergraduates' academic engagement, but the problem applies equally to learning outcomes, retention and

graduation rates, research productivity, and many other measures used to judge institutional performance.

At its core, the issue is that colleges and universities are not monolithic units like cars or toasters but rather are complex composites of many diverse academic disciplines. Each of these disciplines emphasizes distinct modes of instruction, learning, and knowledge generation, and they value demonstrations of those outcomes differently. The consequence is that two different students at the same college can be involved in very different academic activities of the types purported to reflect academic engagement (reading, writing, solving problems, and so on). A growing body of research has documented what is already superficially obvious to any college graduate—that the educational experience of an English major is markedly different than that of an electrical engineering major or economics major. In fact, studies based on two different multi-institutional surveys of student engagement found that *the variation between students in different majors at the same college is greater than the variation between students in the same major at different colleges.*[56] To oversimplify, the broad pattern found in surveys of undergraduates is that humanities and social science majors show the highest level of academic engagement in most areas; engineering, mathematics, and physical science majors show the lowest level; and biological science and professional majors fall somewhere in between. Engineering and science majors, for their part, are more likely to participate in a smaller set of activities such as group work and independent research, and they spend significantly more time in academic pursuits.

Even if they are otherwise valid and reliable, institutional measures combine the scores of individual students from all disciplines and variation in disciplinary composition between

institutions can profoundly influence the institutional comparisons, even to the point that the overall comparison inverts the comparisons of individual disciplines. To illustrate this point, imagine two universities identical in every respect except that University A has 900 seniors in "low-e" majors (major with low engagement as typically measured) and 100 in "high-e" majors, University B has 100 seniors in low-e majors and 900 in high-e majors, and neither has any intermediate-e majors. (Real-life analogues might be the Massachusetts Institute of Technology, with 92 percent low-e majors, and Harvard University, with 9 percent low-e majors, respectively.) In University A, 20 percent (180 out of 900) of low-e majors and 80 percent (80 out of 100) of high-e majors say they discussed academic matters with faculty outside of class. In University B, only 10 percent (10 out of 100) of low-e majors and 40 percent (360 out of 900) of high-e majors did the same. Within each disciplinary group, University A students are twice as likely to have had such discussions, but the 26 percent overall proportion at University A (260 out of 1000) is significantly *lower* than the 37 percent (370 out of 1000) proportion at University B. Prospective students looking at an accountability database would naturally, but mistakenly, conclude that they would be more likely to discuss academic matters with faculty outside of class at University B. This outright reversal of the true underlying statistical relationship, known variously and generally as Simpson's paradox, aggregation bias, or the ecological fallacy, is caused by aggregating outcomes (in this case, to the institutional level) and has potentially dire consequences for comparison of institutional measures.

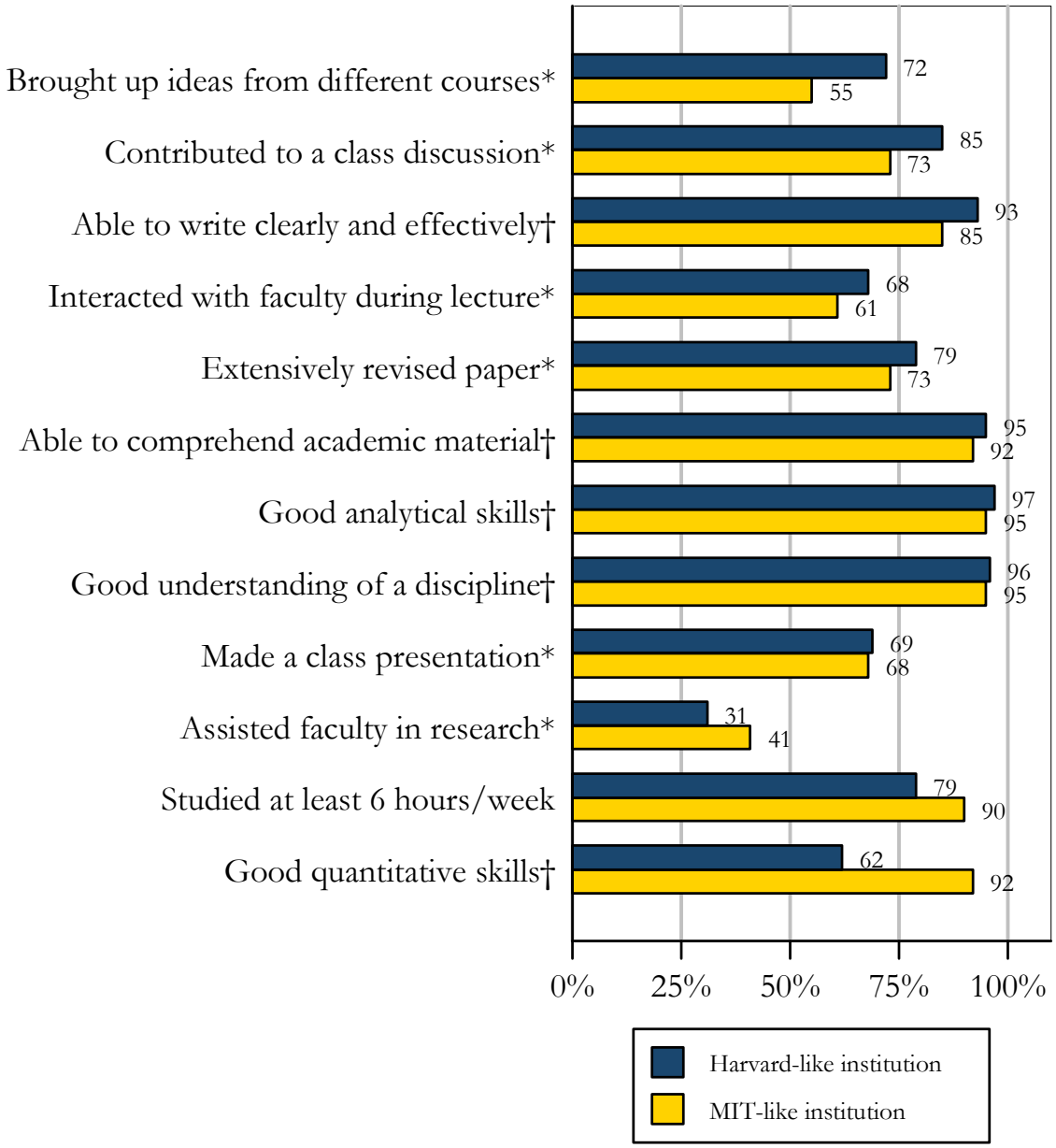
It is difficult to estimate precisely the extent of this effect in actual institutional comparisons, partly because the sample surveys commonly used to assess academic

engagement lack sufficient responses to yield reliable estimates for individual majors. However, the University of California's Student Experience in the Research University census-based SERU/UCUES survey, which in 2008 measured the academic engagement of nearly 60,000 undergraduates, offers fairly fine-grained information about individual disciplines. For example, among seniors in engineering, mathematics, and physical sciences, 72 percent contributed to a class discussion at least occasionally, while 87 percent of seniors in the humanities and social sciences did the same. The 15 percentage point gap is sufficiently large to appreciably alter institutional measures of engagement. Using averages from the UC system, the expected proportion of students contributing to class discussions at MIT (or any university with a similar preponderance of engineering, math, and physical science majors) would be 73 percent and the expected proportion at Harvard (or any similarly liberal-arts intensive university) would be 85 percent, even though the average engagement *within* each discipline is identical. In fact, MIT would be expected to have one of the lowest percentages of undergraduates contributing to class discussions of major research universities nationwide and the second-lowest percentage of any major college or university in Massachusetts (edged out only by Worcester Polytechnic Institute with its mere 1% of its undergraduates in high-e majors). This hypothetical result would be a remarkably poor showing in a state that is arguably the preeminent center of higher education in the country. The impact of this statistical distortion would be even greater if, as recommended by the Spellings report, these figures were used to rank institutions, since college rankings are known to be extremely sensitive to even minute changes.[7]

Figure 1 shows the expected proportion of student responses to a selection of survey-

based attitude and behavioral items typically reported on accountability frameworks. The figures are calculated from a weighted average of high-e and low-e majors for institutions composed similarly to Harvard and MIT using disciplinary averages from the 2008 SERU/UCUES survey. While keeping in mind that these are hypothetical results, several interesting patterns emerge. First, while the Harvard-like university appears to have higher engagement on most measures, there is considerable variation across items, with some items favoring the MIT-like school. At the extremes, the average Harvard-like student would be expected to be 21 percentage points (before rounding) more likely to bring up ideas or concepts from different courses at least occasionally, whereas the average MIT-like student would be 30 percentage points more likely to rate her quantitative skills as good or better. These items were chosen somewhat arbitrarily from a much larger set of survey responses for illustration and not to be strictly representative of any particular accountability framework, but it is not difficult to imagine how deliberately selecting a different group of items could significantly tilt the field toward one type of institution or the other. Moreover, in many cases, such as self-reported analytical and critical thinking skills, frequency of making a class presentation, or attending class at least six hours a week, the difference is a trivial one or two percentage points. Still, the enormous range in differences by disciplinary mix of certain indicators of engagement calls into question the ability to make valid generalizations about the engagement of an entire institution's students.

Figure 1. Selected accountability measures for two hypothetical institutions



*At least "somewhat often"
† Self-reported "good" or better

Source: SERU/UCUES 2008 systemwide results, weighted by major type

If such biased comparisons might merely mislead prospective students, the consequences could be devastating for the institutions themselves. At best, institutional comparisons offer no useful information about which departments are excelling and worthy of emulation and which are lagging and demanding greater attention and resources. At worst, such comparisons could create perverse incentives to boost institutional performance by shuttering departments with the lowest engagement scores whether or not the engagement score was better or worse than expected for the discipline. Such decisions might improve an institution's apparent level of engagement on most indicators and overall, but would anyone claim that MIT would be "better" in any way if it were to become a liberal arts college?

There is no simple comprehensive fix for the problem of institutional measures, but there are better alternatives. For discipline-dependent measures such as student engagement, institutions can report statistics at the level of individual majors or departments. To ensure statistically reliable estimates, institutions will have to expand surveys from samples to the full undergraduate population. This census-based approach has been successfully implemented by SERU since 2004, and starting next year the National Survey of Student Engagement will use this method for freshmen and seniors.

Naturally, reporting statistics by major will increase the length and complexity of accountability reports by an order of magnitude or two. Prospective students and their parents who have a good idea of their intended majors might take the time to browse such comparisons for one or two disciplines, just as consumers shopping for vehicles usually start

with a specific category like minivans or luxury sedans. But not every 17 year-old has decided on a field of study, and it would be the rare high school student (or lawmaker, for that matter) who would compare institutional performance across dozens of different majors. A single set of statistics for each institution, weighted to adjust for disciplinary differences, would be more concise, but it would hide important variation like exceptionally high- or low-performing departments. Unless and until a better solution comes along, the most reasonable response to calls for accountability is for all parties concerned to acknowledge the severe limitations of accountability systems for consumer-oriented information and their susceptibility to bias. The Spellings Report already concedes as much in calling for "value-added" measures of outcomes to account for the vast differences in the college preparedness of students entering different institutions. In light of the obvious difficulties involved, institutional measures of outcomes such as student engagement deserve a similar degree of circumspection so as not to be misinterpreted and misused.

---

[1] U.S. Department of Education. 2006. A Test of Leadership: Charting the Future of U.S. Higher Education. Washington, D.C., pp. 21-22. http://www.ed.gov/about/bdscomm/list/hiedfuture/reports/final-report.pdf

[2] http://voluntarysystem.org

[3] http://www.ucan-network.org

[4] For example, http://www.universityofcalifornia.edu/accountability, http://www.txhighereddata.org/Interactive/Accountability

[5] Nelson Laird, Thomas F., Rick Shoup, and George D. Kuh. 2005. Deep Learning and College Outcomes: Do Fields of Study Differ? Paper presented at the Annual Conference of the California Association for Institutional Research, San Diego.

[6] Chatman, Steve. 2007. Institutional Versus Academic Discipline Measures of Student Experience: A Matter of Relative Validity. Center for Studies in Higher Education, University of California, Berkeley.

[7] Myers, Luke and Jonathan Robe. 2009. College Rankings: History, Criticism and Reform. Center for College Affordability and Productivity, pp. 22-24.

The author thanks Steve Chatman for his many contributions to this paper, including supplying the original inspiration, offering numerous helpful suggestions along the way, and even suggesting the less pejorative term "low-e major." He also gratefully acknowledges the Office of Student Research and Campus Surveys at the University of California, Berkeley, for support of this research. All errors are solely the responsibility of the author.