

The Search for Value-Added: Assessing and Validating Selected Higher Education Outcomes¹

Stephen Klein, RAND; George Kuh, Indiana University; Marc Chun, Council for Aid to Education; Laura Hamilton, RAND; Richard Shavelson, Stanford University

Background and Objectives

The quality of higher education programs has been assessed by (1) examining actuarial data (e.g., graduation rates, minority access, and faculty characteristics); (2) conducting accreditation reviews that consider actuarial data but also inputs, facilities, and school policies; (3) obtaining ratings by faculty and college administrators (such as those published by *US News & World Report*); (4) administering questionnaire surveys that ask students about their collegiate experiences and whether they felt their skills and abilities improved; and (5) testing the students' general education skills. Each method has its advantages and limitations as well as its champions and critics (Klein, 2001; 2002).

While we endorse the use of multiple indicators (e.g., see Riggs and Worthley, 1992; Astin, 1991; Ewell, 1984, 1988; Gentemann et al., 1994; Halpern, 1987; Jacobi et al., 1987; Ratcliff, Jones et al., 1997; Terenzini, 1989; Vandament, 1987), the research reported on here focuses on assessing the technical quality of various open-ended measures of writing and analytic reasoning skills that could be used across colleges. We also are examining the feasibility of using these and other measures to *assess the value that institutions add* (Benjamin & Hersh, 2002) and whether the scores on the open-ended tests are related to a student's admission test scores, college GPA, and responses to a prominent national collegiate survey.

We anticipate that (after adjusting for the students' admissions scores and background characteristics) the new measures will be more

sensitive to program and institutional effects than are traditional tests of general educational abilities (Winter, McClelland and Stewart, 1981; Bohr et al., 1994; Pascarella et al., 1994; Pascarella and Terenzini, 1991).

Measures

To explore these issues, we administered the following measures to students at 14 colleges and universities across the nation (11 in spring 2002 and 3 in fall of 2002):

National Survey of Student Engagement (NSSE). The NSSE project was launched with funding from The Pew Charitable Trusts and is cosponsored by The Carnegie Foundation for the Advancement of Teaching. The instrument has four parts. One part asks students about various experiences they had in college that previous research has found to be related to college grades and other indicators of success, accomplishments, and satisfaction. The second section records student perceptions of key aspects of the institution's environment for learning and the third part asks students to evaluate their own progress.

The last section of the NSSE gathers demographic and other background data on the student (Kuh, 2001). More than 285,000 students at about 620 different four-year colleges and universities have completed the NSSE survey since 2000.

Graduate Record Examination (GRE) essay prompts. The GRE now includes two essay questions. The 45-minute "make-an-argument"

¹Paper presented at the 84th Annual Meeting of the American Educational Research Association, Chicago, April 2003. This research was supported by grants from The Carnegie Corporation, The Ford Foundation, The Hewlett Foundation, ExxonMobil Foundation, and Wabash College Center for Inquiry in the Liberal Arts.

type prompt asks students to justify supporting or not supporting a given position. The 30-minute “break-an-argument” type prompt asks them to critique a position that someone else has taken regarding an issue. Answers can be graded by a trained reader or by a computer (Powers et al., 2000).

Critical Thinking Tests. We used four of the 90-minute “Tasks in Critical Thinking” developed by the New Jersey Department of Education (Ewell, 1994). Each task involves working with various documents and contains several separately scored open-ended questions. We used tasks in science, social science, and arts and humanities.

Performance Tasks. We gave two 90-minute constructed response performance tasks that were modeled after the performance test section of the bar exam (Klein, 1996). Both of these tasks require the student to integrate information from various documents to prepare a memo that provides an objective analysis of a realistic problem.

Task Evaluation Form. This questionnaire asked students about the appropriateness and other characteristics of the constructed response tasks they took. We also conducted focus groups to explore student opinions about the measures and related issues, such as how they could be implemented on an on-going basis on their campuses.

College Transcript. The participants gave their consent for the project to gather data from their college records, including their SAT scores, academic major, college Grade Point Average (GPA), years attending the school, and credit hours earned.

Sample

The study’s 14 colleges and universities varied substantially in size, selectivity, geographic location, and the diversity of their students’ background characteristics (see Table 1). The 1365 students who completed one or more tasks were recruited across academic majors and paid \$20 to \$25 per hour for their participation. There

were similar numbers of freshman, sophomore, junior, and senior participants within a school.

Research Design and Test Administration

At six schools, students were assigned randomly to one of six combinations of measures. Each combination consisted of one GRE make-an-argument essay prompt, one break-an-argument prompt, and either one Critical Thinking task or one Performance Test task. At the other 8 schools, students were assigned randomly to one of 10 combinations of measures. Each combination contained two of the 90-minute measures (one Critical Thinking problem and one Performance Test task or two of one these kinds of measures).

As a result of this matrix sampling, all the 90-minute measures were administered at all schools and all of them were paired with the GRE prompts (but not at all schools). All students also completed the NSSE. A student had the same ID number across measures.

The tests were administered in a controlled setting, usually in one of the school’s computer labs. Students could prepare their answers on a computer, write them long hand, or use a mixture of response modes. The test session took 3 to 3.5 hours, including a short break between measures.

Scaling

To combine results across schools, we used a standard conversion table to put ACT scores on the same scale of measurement as SAT scores and are hereinafter referred to as SAT scores. We converted GPAs within a school to z-scores and then used a regression model (that included the mean SAT score at the student’s school) to adjust all correlations with GPAs for possible differences in grading standards among schools. Finally, to convert the reader assigned “raw” scores on different tasks to a common metric, we scaled the scores on a task to a score distribution that had the same mean and standard deviation as the SAT scores of all the students who took that task.

Score Reliability

We looked at score reliability in several ways. First, table 2 shows that there was a very high degree of agreement between readers. For instance, the mean correlation between two readers was .86 on a GRE prompt and .89 on a 90-minute task.

The mean internal consistency (coefficient alpha) of a 90-minute task was .75, but the mean correlation between any two of them was just .42. The mean correlation between a make and break GRE prompt (.49) was slightly higher. These values (and the .56 correlation between one 90-minute task and a pair of GRE prompts) indicate that the reliability of a total score for a three-hour test battery consisting of two 90-minute tasks or one 90-minute task and two GRE prompts would be about .59 and .71, respectively.

Correlations with Other Measures

SAT scores and GPAs had somewhat higher correlations with scores on a three-hour test battery consisting of both types of GRE prompts and one 90-minute task than they did with a battery containing two 90-minute tasks (Table 3). Part of this difference can be attributed to the differences in the reliability of the scores from these two batteries.

Hand versus machine grading of GRE answers had little or no effect on the correlation of GRE scores with other measures (Table 4). There was a .78 correlation between the hand and machine scoring of the total score across a pair of GRE prompts.²

Regression Analyses

To investigate the unique effect of class, we constructed a regression model where the student was the unit of analysis and the dependent variable was the student's average scale score across all the open-ended tasks that student took. This analysis (which controlled for the student's SAT score, school, and gender) found that the

² Future papers will examine the correlation of our measures with various NSSE scales.

average scores on our measures increased with each class. There was about one quarter of a standard deviation difference between end of spring term freshmen and seniors. These analyses were necessarily restricted to the colleges where testing was done in the spring of 2002.

A regression analysis that controlled for SAT score found that the students who used a computer to draft their answers to the 90-minute tasks earned about one third of a standard deviation higher than did students who hand wrote their answers. Students who used a combination of response modes fell in between these two groups, but were more like those who used the computer.

We also ran regression analyses using the school as the unit of analysis. These analyses (which used all 14 schools) found that a school's mean SAT score explained about 82 percent of the variance in the mean school scores on our outcome measures. Despite this strong correlation and the modest sample sizes, several schools had statistically significantly higher or lower mean scores on our measures (at $p < .05$) than would be expected on the basis of their students' mean SAT scores (see Figure 1).

Student Evaluations of Tasks

An analysis of the Task Evaluation Forms found that 87 percent of the students reported that the time limits were about right or too long. About 65 percent of the students felt the GRE writing prompts were similar to the tasks they had in their college courses whereas 75 percent said the Performance Tasks were mostly different or very different.

Students generally said that the 90-minute tasks we administered were as or more interesting than their usual course assignments and exams, but how much so varied across tasks (see Table 4). The percentage of students rating the overall quality of the measures as good to excellent averaged 69 percent for GRE, 73 percent for Critical Thinking, and 82 percent for the Performance Tasks; but again, ratings varied somewhat by task within type (see Table 5).

Conclusions

Our analyses of the spring and fall of 2002 data indicate that student answers on our critical thinking and writing skills measures can be scored very reliably. This was true for both the 90-minute tasks and the GRE prompts. In addition, the .78 correlation between the hand and computer scoring of a pair of GRE prompts suggests we can rely on computer scoring for school-level analyses. This would result in a significant reduction in scoring time and costs.

Our student level results further suggest that a three-hour test battery consisting of one 90-minute performance task and the two types of GRE prompts would yield total scores that were sufficiently reliable for school-level analyses. This conclusion is consistent with our finding statistically significant school effects after controlling on SAT scores; i.e., despite having only about 100 students per school and SAT scores explaining over 80 percent of the variance in the school level means on our measures.

Implications

This study describes the results with a promising set of cognitive assessment tools that institutions can use to measure general as distinct from domain-specific skills associated with college attendance. These open-ended measures can be administered in a few hours and scored reliably. A machine can even grade some and perhaps eventually all of the answers. This makes these measures very efficient relative to other outcomes assessment batteries. The measures appear to assess important skills that are applicable across major fields. In addition, the scores on them appear to be sensitive to between-institution effects; such findings are rare in the higher education research literature (Pascarella & Terenzini, 1991). These instruments may also prove useful for benchmarking purposes if future studies with larger numbers of institutions replicate our findings of statistically significant between-institution effects.

One of our next steps will be to investigate how scores on our outcomes measures relate to

engagement measures at both the student and school levels. This will allow us to estimate the extent to which those practices that the literature espouses to be educationally effective (Chickering & Gamson, 1987; Kuh, 2001, 2003) translate into cognitive payoffs. With data from enough colleges and universities it also may be possible to develop residual models, whereby we can compare how particular institutions or types of colleges actually score with how they would be predicted to score, given the nature of their students and institutional characteristics. This would open up a potentially instructive approach to measuring institutional effectiveness. In addition, combining these value-added measures with other information about students (e.g., CIRP, NSSE) and institutions will allow us to learn much more about the impact of college on student learning as well as the kinds of educational experiences that contribute to desired college outcomes.

Finally, we found that our measures produced reasonably reliable student-level scores and correlated highly with both SAT scores and GPAs. Moreover, the correlations of our measures with SAT scores and GPAs were both much higher than the relationship between SAT scores and GPAs. If these findings are corroborated by future studies, they may have significant implications for the combination of factors that colleges consider in their admissions process.

References

- Astin, A.W. (1991). *Assessment for excellence*. New York: American Council on Education/Macmillan.
- Benjamin, R. & Hersh, R.H. (2002). Measuring the difference college makes: The RAND/CAE value added assessment initiative. *Peer Review*, 4, 7-10.
- Bohr, L., E. Pascarella, A. Nora, B. Zusman, M. Jacobs, M. Desler, & C. Bulakowski (1994). Cognitive effects of two-year and four-year institutions: A preliminary study. *Community college review*, 22, (1) 4-11.
- Chickering, A. W. & Gamson, Z. F. 1987. Seven principles for good practice in undergraduate education. *AAHE Bulletin*, 39(7), 3-7.
- Ewell, P. T. (1984). *The self-regarding institution: Information for excellence*. Boulder, CO: National Center for Higher Education Management Systems.
- Ewell, P. T. (1988). Outcomes, assessment, and academic improvement: In search of usable knowledge. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research, IV*, 53-108). New York: Agathon Press.
- Ewell, P. T. (1994). *A policy guide for assessment: Making good use of the Tasks in Critical Thinking*. Princeton, NJ: Educational Testing Service.
- Gentemann, K. M., J. J. Fletcher, & D. L. Potter (1994). Refocusing the academic program review on student learning. In M. K. Kinnick (Ed.), *Providing useful information for deans and department chairs* (New Directions for Institutional Research No. 84, 31-46). San Francisco, Jossey-Bass.
- Halpern, D. F. (1987). Recommendations and caveats. In D. F. Halpern (Ed.), *Student outcomes assessment: What institutions stand to gain*. *New directions for higher education*, 59, 109-111.
- Jacobi, M., A. Astin, & F. Ayala (1987). *College student outcomes assessment: A talent development perspective* (ASHE-ERIC Higher Education Report No. 7). Washington, DC: Association for the Study of Higher Education.
- Klein, S. (1996). The costs and benefits of performance testing on the bar examination. *The Bar Examiner*, 65, #3, 13-20.
- Klein, S. (2001). Rationale and plan for assessing higher education outcomes with direct constructed response measures of student skills. New York, NY: Council for Aid to Education, Higher Education Policy Series, Number 3.
- Klein, S. (2002). Direct assessment of cumulative student learning. *Peer Review*, 4, 26-28.
- Kuh, G.D. (2001). Assessing what really matters to student learning: Inside the National Survey of Student Engagement. *Change*, 33(3), 10-17, 66.
- Kuh, G.D. (2003). What we're learning about student engagement from NSSE. *Change*, 35(2), 24-32.
- Pascarella, E. T. & Terenzini, P. T. (1991). *How college affects students: Findings and insights from twenty years of research*. San Francisco: Jossey-Bass.
- Pascarella, E. T., L. Bohr, A. Nora & P.T. Terenzini (1994). *Is differential exposure to college linked to the development of critical thinking?* Illinois Univ., Chicago: National Center on Postsecondary Teaching, Learning, and Assessment.
- Powers, D., Burstein, J.C., Chodorow, M., Fowles, M. E., & Kukich, K. (2000). Comparing the validity of automated and human scoring (GRE Board Report No. 98-08aR). Princeton, NJ: Educational Testing Service.
- Ratcliff, J. L., E. A. Jones, et al. (1997). *Turning results into improvement strategies*. University Park: The Pennsylvania State University, National Center on Postsecondary Teaching, Learning, and Assessment.
- Riggs, M. L. & J. S. Worthley, (1992). Baseline Characteristics of Successful Program of Student Outcomes Assessment, ERIC document ED353285
- Terenzini, P. T. (1989). Assessment with open eyes: Pitfalls in studying student outcomes. *Journal of higher education*, 60, 644-664.
- Vandament, W. E. (1987). A state university perspective on student outcomes assessment. In D. F. Halpern (Ed.), *Student outcomes assessment: What institutions stand to gain*. *New Directions for Higher Education*, 59, 25-28.
- Winter, D. G., D. C. McClelland & A. J. Stewart (1981). *A new case for the liberal arts*. San Francisco: Jossey-Bass.

Table 1. Characteristics of Participating Colleges

School Number	Region	Approx. Enrollment	Type of Funding*	Characteristics
01	Northwest	3,500	Private	Four-year, liberal arts college/average selective admissions/church related
02	Northwest	3,500	Private	Full spectrum teaching/research university/average selective admissions/church related
03	Northwest	6,000	Private	Full spectrum teaching/research university/average selective admissions/ church affiliated
04	Northeast	1,000	Private	Four-year, liberal arts college/highly selective admissions/independent
05	Northeast	2,000	Private	Four-year, liberal arts college/highly selective admissions/independent
06	Northeast	13,000	Private	Independent, full spectrum teaching and research university/non-selective admissions
07	Midwest	1,000	Private	Independent, four-year, single gender, liberal arts college/selective admissions
08	Midwest	1,000	Private	Four-year, liberal arts college/selective admissions/independent
09	Midwest	1,000	Private	Four-year, liberal arts college/selective admissions/church related
10	Midwest	2,000	Private	Four-year, liberal arts college/highly selective admissions/church related
11	Midwest	8,500	Private	Technology oriented research university/ highly selective admissions/independent
12	Midwest	35,000	Public	Full spectrum teaching/research university/selective admissions
13	Southwest	22,000	Public	Full spectrum teaching/research university/non-selective admissions
14	South	6,500	Public	Historic Black university (HBCU)/open admissions

* Public funding also indicates state controlled.

Table 2. Score Reliability

Correlation between readers	
GRE Prompts (hand scoring)	.86
90-minute performance tasks	.89
GRE Reader/Computer correlation	.69
Internal consistency of a 90-minute task	.75
Correlation between	
A make and break GRE prompt	.49
Two 90-minute performance tasks	.42

Table 3. Correlations of a 3-Hour Test Battery with SAT scores and GPA

Tasks in the battery	SAT	GPA
One 90-minute + Two GRE	.69	.64
Two 90-minute task	.47	.51

Table 4. Different GRE Scoring Methods Yield Similar Correlations

Correlation of a pair of GRE prompts with:	GRE Scoring Method	
	Hand	Machine
SAT score	.59	.54
Adjusted GPA	.56	.53
One 90-minute task score	.56	.56

Table 5. Percentage of Students Selecting Each Choice to the Question: “How interesting was this task compared to your usual course assignments and exams?”

Rating Scale	GRE Writing	Critical Thinking				Performance	
		Icarus	Women’s Lives	Conland	Bugs	Drug Policy	Skating
Far more	2	4	7	1	6	3	11
More	18	23	37	22	25	22	43
Average	47	40	41	43	48	44	36
Less	23	21	12	29	18	21	6
Boring	11	12	3	5	3	9	4

Table 6. Percentage of Students Selecting Each Choice to the Question: “What was your *overall* evaluation of the quality of this task?”

Rating Scale	GRE Writing	Critical Thinking				Performance	
		Icarus	Women’s Lives	Conland	Bugs	Drug Policy	Skating
Excellent	2	5	8	3	7	5	8
Very Good	23	19	26	13	24	21	35
Good	44	50	42	51	40	50	45
Fair	25	17	19	28	23	18	10
Poor	4	7	2	4	4	5	1
Very Poor	1	2	1	0	1	1	0
Terrible	0	1	1	1	1	0	0

Figure 1
Plot of Mean School SAT Score by Mean School Overall Task Score

