

1 AN APPROACH TO MEASURING COGNITIVE
2 OUTCOMES ACROSS HIGHER-EDUCATION
3 INSTITUTIONS

4 Stephen P. Klein,*§ George Kuh,** Marc Chun,*** Laura Hamilton,†
5 and Richard Shavelson‡

6
7

8 Over the past decade, state legislatures have experienced increasing pressure to
9 hold higher education accountable for student learning. This pressure stems from
10 several sources, such as increasing costs and decreasing graduation rates. To
11 explore the feasibility of one approach to measuring student learning that
12 emphasizes program improvement, we administered several open-ended tests to
13 1365 students from 14 diverse colleges. The strong correspondence between hand
14 and computer assigned scores indicates the tests can be administered and graded
15 cost eectively on a large scale. The scores were highly reliable, especially when the
16 college is the unit of analysis; they were sensitive to years in college; and they
17 correlated highly with college GPAs. We also found evidence of "value added" in
18 that scores were significantly higher at some schools than at others after controlling
19 on the school's mean SAT score. Finally, the students said the tasks were
20 interesting and engaging.

21
22

23 **KEY WORDS:** value Added; assessment; measuring student outcomes.

24

25

26 Over the past decade, state legislatures have increased the pressure on
27 colleges and universities to become more accountable for student learning.
28 This pressure stems from several sources, including decreasing graduation
29 rates and increasing demand, cost, time-to-degree, economic return and
30 public concern for higher education. Moreover, the federal government

*The RAND Corporation, Santa Monica, California.


**Indiana University, Indiana, IL.

***Council for Aid to Education, New York City, NY.

†The RAND Corporation, Pittsburgh, PA.

‡Graduate Schools of Education and Psychology, Stanford University, Stanford, CA.

§Address correspondence to: Stephen P. Klein, The RAND Corporation, P.O. Box 2138,
Santa Monica, CA 90407-2138, USA. E-mail: stephen_klein@rand.org

	Journal : RIHE	Dispatch : 28-12-2004	Pages : 26
	PIPS No. : NY00001640	<input type="checkbox"/> LE	<input type="checkbox"/> TYPESET
	MS Code : NY00001640	<input checked="" type="checkbox"/> CP	<input checked="" type="checkbox"/> DISK

31 has placed student learning as one of the top priorities for college
32 accrediting agencies.

33 As a result of these concerns, 44 states have created some form of
34 accountability or statistical reporting system (Burke and Minassians,
35 2002) and of those, 27 have formal “report cards” that characterize,
36 among other things, learning outcomes (Naughton, et al., 2003). Of those
37 states that do characterize student outcomes in a report card, Naughton
38 et al. (2003) found 227 performance indicators that were either directly or
39 indirectly related to student learning.

40 Direct indicators of student learning include scores from achievement
41 and ability tests. The most frequent direct indicators are scores on the
42 Graduate Record Examination (GRE), licensure examination pass rates,
43 and infrequently, value-added measures based on published tests (such as
44 CAAP). Indirect indicators are proxies for learning. They typically include
45 graduation rates, degrees awarded, self-reports of learning (such as
46 obtained through the National Student Survey of Engagement), and
47 employer surveys. Data on indirect measures are generally much easier
48 and less expensive to gather than data on direct measures. Consequently,
49 80% of the learning indicators reported by 26 of the 27 states in Naughton
50 et al.’s study (2003) focused on indirect indicators.

51 Increasingly, however, policy debate has focused on direct assessment of
52 student learning (e.g., Callen and Finney, 2002; Klein et al., 2002;
53 Shavelson and Huang, 2003). Not surprisingly, a consensus has not been
54 reached as to what to measure or how to measure it. On the one hand,
55 Callen and Finney have proposed a national assessment for higher
56 education not unlike the National Assessment of Educational Progress
57 (the “Nation’s Report Card” for K-12 education). Such an assessment
58 would permit state-by-state comparison for policy makers that would be
59 reported in “*Measuring Up*,” a biennial higher education report card.
60 Such an approach provides information to state policy makers for
61 decision-making but this level of aggregation is not particularly useful for
62 institutional improvement. In contrast, some researchers (e.g., Benjamin
63 and Hersh, 2002; Klein, 2001; Klein et al., 2003) have proposed a multi-
64 level assessment adapted to local institutions’ concerns for the improve-
65 ment of teaching and learning. In such a system, comparison sets of
66 cooperating institutions could participate and benchmark progress.
67 Results of such an assessment system could also be included in state
68 report cards.

69 In our view, the latter approach is more likely to lead to improved
70 learning in America’s diverse institutions of higher education than the
71 former. However, we cannot at present test this assumption without
72 alternatives to compare. In the long run we believe that a rapprochement is

73 needed between the legitimate demands of policy makers and the calls for
74 institutional improvement. This paper is a step in that direction. It reports
75 the results of an effort to assess important aspects of student learning in
76 higher education. Our approach integrates cognitive outcomes associated
77 with learning with thinking and writing ability outcomes associated with
78 broader goals of undergraduate general education. The findings represent
79 direct measures of these broader goals using the college rather than the
80 individual student (or the state) as the unit of analysis. Specifically, the
81 study addresses the following questions:

- 82 1. Can we obtain reasonably reliable school level scores on open-ended (as
83 distinct from multiple choice) tests of the students' writing and critical
84 thinking skills?
- 85 2. After controlling for "input" (as measured by SAT or ACT scores), do
86 the students' scores improve as they progress through the college years?
- 87 3. Are the students' scores on our measures related to their GPAs? In
88 other words, are we measuring skills and abilities that appear to in-
89 fluence and/or reflect learning?
- 90 4. After controlling for "input" do the students at some colleges generally
91 earn statistically significantly higher scores than students at other col-
92 leges? That is, are the measures sensitive to the "value added" by the
93 institution?
- 94 5. Do students find the measures interesting and engaging enough to
95 motivate them to try their best?
- 96 6. Can the measures be administered (and the answers to them scored) in
97 an efficient, timely, and cost effective way? For example, is it feasible to
98 use computer scoring of essay answers on a large scale?

99 The next section of this paper summarizes the approaches that have
100 been used in the past to assess the effectiveness of educational programs.
101 We follow this with a review of the conceptual framework Shavelson and
102 Huang (2003) suggested for defining and building direct measures of
103 student learning. The remainder of this article presents our findings
104 regarding the questions above and discusses their implications.

105 BACKGROUND: PAST EFFORTS TO MEASURE 106 STUDENT LEARNING

107 To provide a context for what follows, we briefly review past efforts to
108 assess institutional quality including student cognitive outcomes. These
109 efforts have generally relied on one or more of the following four methods:
110 (1) tabulating actuarial data; (2) obtaining ratings of institutional quality;

111 (3) conducting student surveys; and (4) directly measuring student skills
112 and knowledge.

113 Actuarial Data

114 Colleges routinely report various types of actuarial data, such as
115 graduation rates, endowment level, student/faculty ratio, average admis-
116 sions test scores, and the racial/ethnic composition of the student body.
117 The advantages of such indices are that the data for them are relatively
118 straightforward to collect and the resulting statistics can be compared
119 over time and (with caution) across institutions. Although not intrinsic to
120 the data themselves, the way in which the analyses are conducted
121 typically assumes that a better quality educational institution (or a better
122 quality educational experience) is associated with more and better
123 resources—in this case, better funding, better faculty (which is defined
124 as a higher percentage of any given cadre holding Ph.D.s), and better
125 students as reflected by higher admissions selectivity (Astin, 1968, 1977,
126 1991, 1993).

127 Actuarial data have been used by some states to measure institutional
128 effectiveness (Gates et al., 2001). They also have been used by the National
129 Center for Education Statistics (NCES) and the Integrated Postsecondary
130 Education Data System (IPEDS), which include data on student
131 enrollment, faculty ranks, and institutional expenditures. These national
132 databases are large in scope, with some of the data coming from secondary
133 sources—such as census counts and transcripts (NCHEMS, 1994).
134 Reviews of national data systems suggest that they yield little information
135 about an institution's effectiveness in promoting student cognitive
136 outcomes (Dey et al., 1997; National Postsecondary Education Coopera-
137 tive, 2000a, b).

138 Ratings of Institutional Quality

139 Ratings of institutional quality are generated annually from surveys of
140 college faculty and administrators, but also may include actuarial data such
141 as selectivity, faculty resources, and financial resources. Although using
142 multiple indicators and measures is consistent with the good assessment
143 practice (e.g., see Astin, 1991; Ewell, 1984, 1988; Gentemann, Fletcher and
144 Potter 1994; Halpern, 1987; Jacobi, Austin and Ayala, 1987; Ratcliff et al.,
145 1997; Riggs and Worthley, 1992; Terenzini, 1989; Vandament, 1987),
146 college rankings (such as those produced by the *U.S. News and World*
147 *Report*) have come under heavy fire, including from highly-rated
148 institutions. For example, a 1997 report by the National Opinion Research

149 Center (commissioned by *U.S. News and World Report*) was highly critical
150 of the weighting scheme, the subjective nature of the ratings, and the role of
151 reputations in the ranking. Additional problems have been noted by others
152 (see Graham and Thompson, 2001; Klein and Hamilton, 1998; Machung,
153 1995; McGuire, 1995; Winter, McClelland and Stewart, 1981).

154 Student Surveys

155 Large-scale questionnaire surveys have been used to ask students about
156 their collegiate experiences, satisfaction with their coursework and school,
157 self-assessments of improvement in their academic abilities, and educa-
158 tional and employment plans (Astin, 1991; Ewell, 1987; Gill, 1993;
159 Johnson et al., 1993; Lenning, 1988; Muffo and Bunda, 1993). Interviews
160 of individuals or groups also have been used (Johnson et al., 1993;
161 Lenning, 1988; Smith et al., 1993). The main advantage of these surveys is
162 that they can gather a large amount of data economically about an
163 institution (NCHEMS, 1994). Survey results also have been used to assess
164 and compare institutional effectiveness (Astin, 1993; Pace, 1990; Terenzini
165 and Wright, 1987).

166 There are three prominent examples of this approach. The Baccalaureate
167 and Beyond Longitudinal Study, based on the National Postsecondary
168 Student Aid Study, gathers information about education and work
169 experiences after student completion of the bachelor's degree. The
170 Cooperative Institutional Research Program (CIRP) survey, administered
171 by UCLA's Higher Education Research Institute (HERI), asks entering
172 freshmen to report on activities, goals, and self. The National Survey of
173 Student Engagement (NSSE) carries on and extends this tradition by
174 asking questions about features of college life that previous research has
175 found to be associated with improved student performance (Kuh, 2001).
176 The limitation of these questionnaires for assessing student outcomes are
177 inherent to any survey, such as whether students can accurately report
178 how much their college experiences have improved their analytic and
179 critical thinking skills.

180 Direct Assessments of Student Learning

181 A fourth approach to assessing the quality of an institution's
182 educational programs measures student learning directly (Winter, McClel-
183 land, and Stewart, 1981). Direct assessments may involve collecting data
184 on course grades, evaluating student work products (e.g., portfolios), and
185 administering various types of tests. An institution's faculty and staff
186 typically conduct these efforts on their own students, although some

187 institutions have collaborated in using the same measures to assess
188 learning outcomes. The latter strategy allows institutions and policy
189 makers to compare institutions (Obler, Slark and Umbdenstock, 1993;
190 Bohr et al., 1994; and Pascarella et al., 1996). A few states have required
191 that all institutions use the same standardized multiple-choice tests to
192 assess student knowledge, skills, and abilities (Cole, Nettles and Sharp,
193 1997; Naughton, Suen and Shavelson, 2003; NCHEMS, 1996; Steele and
194 Lutz, 1995). These methods have been used to collect data on individual
195 students and on groups of students, at the program and at the institutional
196 level (Ratcliff et al., 1991).

197 In addition to the more commonly used paper and pencil examinations,
198 direct assessments of students include portfolios (Banta et al., 1996; Black,
199 1993; Fong, 1988; Forrest, 1990; Hutchings, 1989; Johnson et al., 1993;
200 Suen and Parkes, 1996; Waluconis, 1993) and on-demand performances,
201 such as presentations, debates, dances, and musical recitals (Palomba and
202 Banta, 1999). Researchers disagree about the validity of such approaches.
203 One such concern is the lack of standardization across tasks, another is the
204 question of who actually did the work (which is not a problem for a
205 student giving a recital but is an issue for term papers and other work
206 products created outside of class), and still another is score reliability
207 when the results are used to make decisions about individual students;
208 although this may not be a major problem if the data are used to assess
209 program effects (Johnson et al., 1993; Lenning, 1988).

210 Course grades are an obvious choice as an outcome measure, but they
211 are specific to individual professors. Course grades, then, are difficult to
212 compare even across faculty within a school. They are even more difficult
213 to compare across colleges because of large differences in admissions and
214 grading standards. Finally, the current debate over grade inflation
215 highlights the problem of using grades over time to monitor progress.

216 The shortcomings of these measures as indicators of learning outcomes
217 have led some to suggest comparing colleges on how well their students do
218 on graduate and professional school admissions tests and licensing exams
219 (such as for teachers, accountants, and engineers). However, this approach
220 is fraught with problems, such as concerns about the relevance of these
221 measures to the goals of undergraduate programs, selection bias in who
222 prepares for and takes these tests, and student motivation and related
223 issues (which contributed to the Air Force Academy discontinuing its
224 requirement that all of its graduating seniors take the GREs).

225 Our approach to direct assessment takes a different tack. The tasks are
226 all open-ended (rather than multiple-choice) and matrix-sampled within a
227 college (i.e., each student takes only a few of the several different tests
228 used) so that a wide range of tasks can be administered. This is done to

229 reduce the response burden on individual students while still allowing
230 coverage of a broad spectrum of areas.

231 CONCEPTUAL FRAMEWORK

232 The measures we are using fit within Shavelson and Huang's (2003)
233 framework for conceptualizing, developing, and interpreting direct
234 measures of students' learning. This framework utilized past research on
235 cognition and human abilities (e.g., Gustafsson and Undheim, 1996;
236 Martinez, 2000; Messick, 1984; Pellegrino, Chudowsky and Glaser, 2001)
237 to characterize alternative ways of measuring college students' knowledge
238 and learning.

239 There are at least three reasons why the Shavelson and Huang
240 framework is useful for assessing higher education learning outcomes.
241 First, and most importantly, it clarifies the constructs we are and are not
242 measuring and the higher education goals associated with them. The
243 framework, then, guides instrument construction/selection and interpreta-
244 tion. Second, the framework shows where our constructs fit within a
245 100-year history of efforts to assess student learning in higher education
246 and what has been measured in the past. Third, some of the visions of
247 student learning being proposed by others for higher education initially
248 appear to be inconsistent and contradictory. The framework allows us to
249 integrate and represent these visions.

250 Following Shavelson and Huang, cognitive outcomes in higher
251 education range from domain-specific knowledge acquisition to the most
252 general of reasoning and problem-solving abilities, to what Spearman
253 called general ability or simply "G." (We refer to "G" to avoid antiquated
254 interpretation of g as genetically determined; see Cronbach, 2000;
255 Kyllonen and Shute, 1989; Messick, 1984; Snow and Lohman, 1989).
256 Yet we know that learning is highly situated and context bound.¹ Only
257 through extensive engagement, practice and feedback in a domain does
258 this knowledge, interacting with prior knowledge and experience, become
259 increasingly decontextualized so that it transfers to enhance general
260 reasoning, problem solving and decision making in a broad domain and
261 later to multiple domains (e.g., Bransford, et al., 1999; Messick, 1984).

262 What is learned (and to what level it transfers) depends on the aptitudes
263 and abilities that students bring with them from their prior education (in
264 and out of school) and their natural endowments (e.g., Shavelson et al.,
265 2002). A useful framework for linking outcomes with assessments, then,
266 must capture this recursive complexity. It must allow us to map the
267 proposed tests onto the knowledge and abilities that are so highly valued
268 as cognitive outcomes in higher education.

269 As shown in Table 1, levels I–VI in the Shavelson and Huang
 270 framework move from “abstract/process oriented” at the top of the table
 271 to “concrete content oriented” abilities at the bottom. This ordering also
 272 corresponds to abilities that are based on “inheritance interacting with
 273 accumulated experience” to those based on “direct experience.”

274 General abilities, such as verbal, quantitative and visual-spatial
 275 reasoning (see Carroll, 1993), build on inherited capacities and typically
 276 develop over many years in formal and informal education settings. These
 277 abilities contribute to fluid intelligence (closely allied with “G” and
 278 indirectly related to prior learning from a wide range of experiences) and
 279 crystallized intelligence (closely allied with learning experiences). “[F]luid
 280 intelligence is functionally manifest in novel situations in which prior
 281 experience does not provide sufficient direction, crystallized intelligence is
 282 the precipitate of prior experience and represents the massive contribution
 283 of culture to the intellect” (Martinez, 2000, p. 19). However, measures of
 284 crystallized, fluid, and general intelligence do not adequately reflect the in-
 285 college learning opportunities available to students. They are included in
 286 Table 1 for completeness only.

287 Shavelson and Huang acknowledged that their hierarchy oversimplifies.
 288 Knowledge and abilities are interdependent. Learning depends not only
 289 on instruction but also on the knowledge and abilities students bring to
 290 college. Indeed, instruction and abilities are likely to interact to produce
 291 learning, and the course of this interaction evolves over time so that
 292 different abilities are called forth and different learning tasks are needed in
 293 this evolution (Shavelson et al., 2002; Snow, 1994). Thus, Table 1 does not

**TABLE 1. Location of the Study’s Measures in Shavelson/Huang’s
 Conceptual Framework**

Level	What is Measured	Measures Used
I	General intelligence (“G”)	
II	Fluid and crystallized intelligence	
III	Verbal, quantitative, and spatial reasoning	SAT-I scores
IV	Reasoning, comprehending, problem solving, and decision making <i>across</i> broad domains (humanities, social sciences, sciences)	GRE writing prompts: make and break an argument
V	Reasoning, comprehending, problem solving, and decision making <i>within</i> broad domains (humanities, social sciences, sciences)	Performance and critical thinking tasks
VI	Declarative, procedural, schematic, and strategic domain-specific knowledge	College GPA

Note: A measure may overlap more than one level in the framework.

294 behave in strict hierarchical fashion. It is intended to be heuristic, to
295 provide a conceptual framework for discussing and developing learning
296 measures. The research reported here focuses on levels III–VI of this
297 framework and especially on the cusp between III and IV.

298 By *domain-specific knowledge*, Shavelson and Huang were referring to
299 knowledge of specific subjects, such as chemistry or history. This is the
300 kind of knowledge we would expect to see assessed in students' learning
301 within an academic major. Domain-specific knowledge corresponds to
302 such valued outcomes of higher education (goals) as are typically labeled,
303 "learning high-tech skills" or "specific expertise and knowledge in chosen
304 career." Shavelson and Huang (2003) divided domain-specific knowledge
305 into the following four types: declarative ("knowing that"), procedural
306 ("knowing how"), schematic ("knowing why"), and strategic ("knowing
307 when, where and how"—knowing when certain knowledge applies, where
308 it applies, and how it applies).

309 Tests of domain knowledge are appropriate measures of student
310 learning in a major and should be included in the assessment of student
311 learning. Such tests may be published, such as the GRE's area tests. Yet
312 the GRE tests are no longer widely used in most academic majors for a
313 number of reasons including, among others, their fit with the department's
314 particular definition of the major. We also know that students' knowledge
315 in their academic majors is tested extensively by individual instructors
316 and, in some cases, in a capstone course or by an integrated examination.
317 We believe that capitalizing on the availability of such tests provides an
318 opportunity to assess domain-specific knowledge in context. We plan to
319 explore some simple, straightforward ways of doing this, such as by using
320 a pretest, intermediary, and final exams in core (or capstone) courses in the
321 major to examine gain-score effect sizes with and without adjusting for
322 SAT (or ACT) scores.

323 *Broad abilities* are complexes of cognitive processes ("thinking") that
324 underlie verbal, quantitative and spatial reasoning, comprehending,
325 problem solving and decision making in a domain, and more generally
326 across domains. These abilities are developed well into adulthood through
327 learning in and transfer from non-school as well as school experiences,
328 repeated exercise of domain-specific knowledge in conjunction with prior
329 learning and previously established general reasoning abilities. As the
330 tasks become increasingly broad—moving from a knowledge domain to a
331 field such as social science, to broad everyday problems—general abilities
332 exercise greater influence over performance than do knowledge structures
333 and domain-specific abilities. Many of the valued outcomes of higher
334 education are associated with the development of these broad abilities.
335 For example, two important goals identified in the National Center for

336 Public Policy and Higher Education's (Immerwahl, 2000) survey were
337 "improved problem solving and thinking ability," and "top-notch writing
338 and speaking."

339 Assessments of learning currently in vogue, as well as some assessments
340 developed in the mid-20th century, tap into these broad abilities. Most have
341 focused primarily at the level of the sciences, social sciences, and humanities.
342 The science area score falls between domain specific knowledge and general
343 reasoning abilities. Other tests are more generic, focusing on critical writing
344 and reasoning. Some examples are the GRE's Analytic writing prompts,
345 the College-BASE, the Academic Profile, CAAP, UAP Field Tests, and the
346 90-minute tasks used in this study. Indeed, many tests of broad abilities
347 contain both area (e.g., sciences) and general reasoning and writing tests.

348 ACT and the Educational Testing Service have both offered "general
349 education" measures in reading, writing, and mathematics, such as for
350 "rising junior" exams. While few would argue against college students
351 being proficient in these areas, there is little evidence that scores on such
352 measures are sensitive to the effects of different types of college level
353 programs. For example, 23 institutions participated in perhaps the most
354 comprehensive *longitudinal* study of learning at the college level to date
355 (Pascarella et al., 1996). This study "found little evidence to suggest that
356 attending an academically selective four-year institution had much impact
357 on growth in critical-thinking skills during the first three years of college"
358 (Pascarella, 2001, p. 22).

359 A *cross-sectional* study at 56 institutions found that most of the
360 improvement in skills occurs in the first two years of college (Flowers,
361 et al., 2001). However, both the Pascarella et al. and the Flowers et al. studies
362 relied on multiple-choice tests of general education skills. There were no open-
363 ended measures (even in writing) and the tests they used did not ask students
364 to apply their abilities to realistic and inherently engaging complex tasks.

365 INVESTIGATING THE FEASIBILITY OF USING 366 DIRECT MEASURES

367 This section describes the procedures we used and the results obtained
368 in our initial exploration of the feasibility and utility of using open-ended
369 direct measures of student learning. Specifically, we examined whether
370 measures that were designed to be more aligned with the abilities that
371 colleges say they are trying to develop (and focused on levels III through V
372 in the conceptual framework) could be administered efficiently, whether
373 the responses to these tasks could be scored consistently, whether the
374 scores on them were reliable enough across tasks to have confidence in the
375 school level results, whether those scores were related to years in college

376 and possible differences in programs across institutions, and whether the
377 tasks were engaging enough to motivate students to try their best on them.
378 We also discuss the implications of our findings.

379 Participating Institutions

380 Presentations by project staff at professional conferences led to over two-
381 dozen colleges and universities offering to participate in the study. We
382 selected 14 of these schools so that as a group they had a very diverse set of
383 characteristics, including geographical location, size, primary funding source
384 (public versus private), admissions selectivity, and Carnegie classification (see
385 Table 2 for details). No attempt was made to draw a random sample of the
386 higher education sector, but rather to reflect its diversity given that our goal
387 was to examine the feasibility of using the measures rather than reporting
388 normative results or data about individual institutions.

389 Sampling Students Within Colleges

390 The 1365 students who participated in this research were recruited
391 across academic majors and paid \$20 to \$25 per hour for their
392 participation (the amount varied as function of local practices and
393 policies). Colleges were asked to select about equal numbers of freshman,
394 sophomores, juniors, and seniors so that all told there would be about 100
395 students per school. Recruitment methods varied. For example, some
396 schools offered participation to all students and then took the first 25–30
397 who applied in each class whereas others used a sophisticated stratified
398 random sampling procedure. Participation was optional at all campuses.
399 Thus, it is not appropriate to report or compare individual school means
400 (nor was it ever our intention or need to do so).

401 Measures

402 The GRE prompts described below were used at six colleges. All of the
403 other measures were used at all 14 colleges. Table 1 shows the location of
404 each of the cognitive measures in the Shavelson/Huang framework.

405 *Graduate Record Examination (GRE) Essay Prompts*

406 The GRE now includes two essay questions. The 45-minute “make-an-
407 argument” type prompt asks students to justify supporting or not
408 supporting a given position. The 30-minute “break-an-argument” type
409 prompt asks them to critique a position that someone else has taken
410 regarding an issue (see Powers et al., 2000 for examples).

TABLE 2. Characteristics of Participating Colleges

School Number	Region	Approx. Enrollment	Type of Funding ^a	Characteristics
01	Northwest	3500	Private	Four-year, liberal arts college/average selective admissions/church related
02	Northwest	3500	Private	Full spectrum teaching/research university/average selective admissions/church related
03	Northwest	6000	Private	Full spectrum teaching/research university/average selective admissions/church affiliated
04	Northeast	1000	Private	Four-year, liberal arts college/highly selective admissions/independent
05	Northeast	2000	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	1000	Private	Independent, four-year, single gender, liberal arts college/selective admissions
08	Midwest	1000	Private	Four-year, liberal arts college/selective admissions/independent
09	Midwest	1000	Private	Four-year, liberal arts college/selective admissions/church related
10	Midwest	2000	Private	Four-year, liberal arts college/highly selective admissions/church related
11	Midwest	8500	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	6500	Public	Historic Black university (HBCU)/ open admissions

^aPublic funding also indicates state controlled.

411 *Critical Thinking Tests*

412 We used four of the 90-minute “Tasks in Critical Thinking” developed
 413 by the New Jersey Department of Education (Ewell, 1994). Each task

414 involves working with various documents and contains several separately
415 scored open-ended questions. We used tasks in science, social science, and
416 arts and humanities.

417 *Performance Tasks*

418 We developed and administered two 90-minute constructed response
419 tasks that were modeled after the performance test section of the bar exam
420 (Klein, 1996). These tasks require students to integrate information from
421 various documents to prepare a memo that provides an objective analysis
422 of a realistic problem (see Klein et al., 2004 for an example).

423 *Task Evaluation Form*

424 This questionnaire asked students about the appropriateness and other
425 characteristics of the tests they took. We also conducted focus groups to
426 explore student opinions about the measures and related issues, such as
427 how they could be implemented on an on-going basis on their campuses.

428 *College Transcript*

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

433 *National Survey of Student Engagement (NSSE)*

434 The NSSE has four parts. One part asks students about experiences they
435 had in college that previous research has found to be related to college
436 grades and other indicators of success, accomplishments, and satisfaction.
437 The second section records students' perceptions of key aspects of the
438 institution's environment for learning and the third part asks students to
439 evaluate their own progress. The last section of the NSSE gathers
440 demographic and other background data on the student (Kuh, 2001).
441 More than 430,000 students at about 730 different four-year colleges and
442 universities have completed the NSSE survey since 2000.

443 **Research Design and Test Administration**

444 At six schools, students were assigned randomly to one of six
445 combinations of measures. Each of these combinations consisted of one

TABLE 3. Matrix Sampling Plan at the Six Group 1 Colleges

Set	90-Minute Task	GRE Tasks	
		Choice	No Choice
1	Icarus Myth	A1 or A2	B3
2	Women's Lives	A1 or A3	B1
3	Conland & Teresia	A2 or A3	B2
4	Mosquitoes	A1 or A3	B2
5	Crime Reduction	A1 or A2	B1
6	SportsCo	A2 or A3	B3

Note: There were three GRE “make-an-argument” prompts (A1, A2, and A3). Students were given two of them and instructed to pick one to answer. They also were given one of three break-an-argument prompts. Sets were assigned randomly to students within a school.

446 GRE make-an-argument essay prompt, one break-an-argument prompt,
 447 and either one Critical Thinking task or one Performance Test task (see
 448 Table 3 for details). At the other 8 schools, students were assigned
 449 randomly to one of 10 combinations of measures. Each of these
 450 combinations contained two of the 90-minute measures (see Table 4 for
 451 details).

452 All of the tests were administered in a controlled setting, usually in one
 453 of the college's computer labs. Students could prepare their answers on a
 454 computer, write them long hand, or use a mixture of response modes. The
 455 test session took 3–3.5 hours, including a short break between measures.

TABLE 4. Matrix Sampling Plan at the Eight Group 2 Colleges

Set	First 90-Minute Task	Second 90-Minute Task
7	Conland & Teresia	Women's Lives
8	Mosquitoes	Icarus Myth
9	Woman's Lives	Mosquitoes
10	Icarus Myth	Conland & Teresia
11	Crime Reduction	Icarus Myth
12	SportsCo	Woman's Lives
13	Conland & Teresia	Crime Reduction
14	Mosquitoes	SportsCo
15	SportsCo	Crime Reduction
16	Crime Reduction	SportsCo

Note: Each student took two 90-minute tasks with a 5-minute break between them. Sets were assigned randomly to students within school.

456 Testing was conducted in the spring of 2002 at 11 of the 14 colleges and in
457 the fall of 2002 at the other three schools.

458 Scaling

459 To combine results across colleges, we used a standard conversion table
460 to put ACT scores on the same scale of measurement as SAT scores and
461 are hereinafter referred to as SAT scores. We converted GPAs within a
462 school to *z*-scores and then used a regression model (that included the
463 mean SAT score at the student's college) to adjust the correlations with
464 GPAs for possible differences in grading standards among colleges.
465 Finally, to convert the reader assigned "raw" scores on different tasks to a
466 common metric, we scaled the scores on a task to a score distribution that
467 had the same mean and standard deviation as the SAT scores of all the
468 students who took that task.

469 Scoring

470 Answers to the GRE prompts and the four critical thinking tasks were
471 graded by a two-person team that had extensive experience in scoring
472 answers to these prompts. A four-person team graded the answers to the
473 two 90-minute performance tasks. The answers to the GRE prompts and
474 SportsCo also were machine scored (see Klein et al., 2004 for details).
475 Except where noted otherwise, the results below are based on the hand
476 (rather than the machine) assigned scores.

477 Score Reliability

478 Table 5 shows that there was a very high degree of consistency between
479 readers. For instance, the mean correlation between two readers was .86
480 on a GRE prompt and .89 on a 90-minute task. The mean internal
481 consistency (coefficient alpha) of a 90-minute task was .75; the mean
482 correlation between any two of them was .42. The mean correlation
483 between a make and break GRE prompt (.49) was slightly higher. These
484 values (and the .56 correlation between one 90-minute task and a pair of
485 GRE prompts) indicate that the reliability of an individual student's total
486 score for a 3-hour test battery consisting of two 90-minute tasks or one 90-
487 minute task and two GRE prompts would be about .59 and .71,
488 respectively. The reliability of a college's mean score would exceed .90.

489 Response Mode Effects

490 A regression analysis that predicted a student's score on our 90-minute
491 measures on the basis of that student's SAT score and response mode

TABLE 5. Indices of Score Reliability (Medians)

Correlation between Two Readers on a Single GRE Prompt (hand scoring)	.86
90-Minute task	.89
Correlation between reader and computer assigned scores on a single GRE prompt	.69
Internal consistency of a 90-minute task	.75
Correlation between Hand scored make and break GRE prompts	.49
Two 90-minute tasks	.42

Note: The internal consistency (coefficient alpha) of a 90-minute task was based on the correlations among the separately scored questions or issues in that task.

492 (i.e., handwrite and/or use a computer to draft answers) found that the
493 students who used a computer earned about one third of a standard
494 deviation higher score than did students who hand wrote their answers.
495 Students who used a combination of response modes fell in between these
496 two groups, but were more like those who used the computer.

497 Correlations with Other Measures

498 Student SAT scores and GPAs had somewhat higher correlations with
499 scores on a 3-hour test battery consisting of both types of GRE prompts
500 and one 90-minute task than they did with a battery containing two
501 90-minute tasks (Table 6). Some of this difference can be attributed to the
502 differences in the reliability of the scores from these two batteries.²

503 Hand versus Machine Scoring

504 At the individual student level, there was a .78 correlation between the
505 hand and machine scoring of the total score across a pair of GRE prompts.
506 We also found that computer grading of the answers to a

TABLE 6. Correlations of a Three-Hour Test Battery with SAT Scores and GPA

Tasks in the Battery	SAT Scores	Adjusted GPA
One 90-minute + two GRE	.69	.64
Two 90-minute tasks	.47	.51

Note: The correlation between SAT scores and adjusted GPA was .60. The *within* school correlation between SAT scores and GPA ranged from .12 to .60 with a median of .37.

507 90-minute task produced scores that were very comparable to those
 508 assigned by hand. For instance, at the individual student level, there was
 509 a .84 correlation between the hand and computer scores on the SportsCo
 510 task. This correlation increased to .95 when the school is used as the unit of
 511 analysis. The method used to grade the answers (i.e., hand versus machine)
 512 had little or no effect on the correlation of the resulting scores with other
 513 measures, such as SAT scores (see Table 7 and Klein et al., 2004).

514 Changes in Performance from Freshman to Senior Year

515 To explore whether our open-ended measures were sensitive to changes
 516 in student ability over time (i.e., from freshman to senior year), we
 517 constructed a regression model where the student was the unit of analysis
 518 and the dependent variable was the student's average scale score across all
 519 the tasks that the student took. This analysis (which controlled for the
 520 student's SAT score, school, and gender) found that the average scores on
 521 our measures increased with each class. Specifically, there was about one
 522 quarter of a standard deviation difference between end-of-spring-term
 523 freshmen and seniors. These analyses were necessarily restricted to the 11
 524 colleges where testing was done in the spring of 2002.

525 School Effects

526 To study the potential impact of college on student achievement, we
 527 regressed the mean score for each college on its average SAT score. This
 528 equation also had a dummy variable for each college. This analysis found
 529 that mean SAT scores by themselves explained about 82% of the variance
 530 in mean college achievement scores. Despite this strong correlation and
 531 the modest sample sizes, three colleges had statistically significantly higher
 532 or lower mean scores on our measures (at $p < .05$) than would be expected
 533 on the basis of their students' mean SAT scores. Figure 1 shows this
 534 relationship (there is one data point plotted for each school).

TABLE 7. Different GRE Scoring Methods Yield Similar Correlations

Correlation of a Pair of GRE Prompts with:	GRE Scoring Method	
	Hand	Computer
SAT score	.59	.54
Adjusted GPA	.56	.53
One 90-minute task score	.56	.56

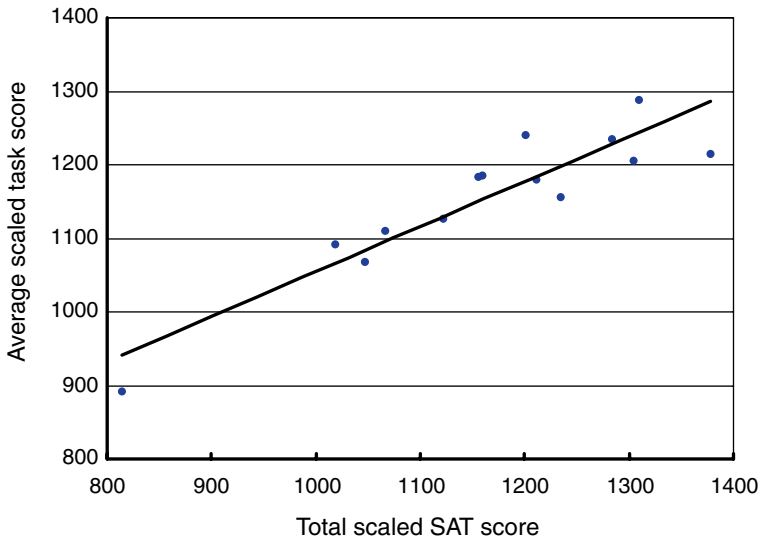


FIG. 1. Relationship between a College's mean SAT and scaled task score.

535 Student Evaluations of Tasks

536 An analysis of the Task Evaluation Forms found that 87% of the
 537 students reported that the time limits were about right or too long. About
 538 65% of the students felt the GRE writing prompts were similar to the
 539 tasks they had in their college courses whereas 75% said the Performance
 540 Tasks were mostly different or very different.

541 Students generally said that the 90-minute tasks were as much or more
 542 interesting than their usual course assignments and exams, but how much
 543 so varied across tasks (see Table 8). The percentage of students rating the
 544 overall quality of the measures as good to excellent averaged 69% for
 545 GRE, 73% for Critical Thinking, and 82% for the Performance Tasks;
 546 but again, ratings varied by task within type (see Table 9).

547 CONCLUSIONS

548 To sum up, we examined a set of constructed-response tests that were
 549 designed to tap student analytic reasoning and writing skills. Our goal was
 550 to examine the utility of these measures when the institution rather than
 551 the student is the primary unit of analysis.

552 We found that student answers on our measures can be scored very
 553 reliably. This was true for both the 90-minute tasks and the GRE prompts.

TABLE 8. 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	90-Minute Critical Thinking Performance Tasks					
		Icarus Myth	Women’s Lives	Conland & Teresia	Mosquitoes	Crime	SportsCo
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 9. Percentage of Students Selecting Each Choice to the Question: “What was Your Overall Evaluation of the Quality of This Task?”

Rating Scale	GRE Writing	90-Minute Critical Thinking Performance Tasks					
		Icarus Myth	Women’s Lives	Conland & Teresia	Mosquitoes	Crime	SportsCo
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

554 The .78 correlation between the hand and computer scoring of a pair of a
 555 student’s answers to the GRE prompts and the .84 correlation between
 556 these two scoring methods on a 90-minute task were impressive. More
 557 importantly, the near perfect correlation between these scoring methods
 558 when the school is used as the unit of analysis suggests that in the future
 559 we can rely on machine scoring for school-level analyses. This would result
 560 in a significant reduction in scoring time and costs. For example, the cost
 561 for one person to grade the response to a single GRE writing prompt is
 562 about \$2.50, but about \$1.00 for the computer grading. The computer also
 563 is much faster for reporting scores.

564 Our results further suggest that a 3-hour test battery consisting of one
 565 90-minute performance task and the two types of GRE prompts would
 566 yield total scores that were sufficiently reliable for school-level analyses.
 567 This conclusion is consistent with our finding statistically significant

568 school effects after controlling on SAT scores; i.e., despite having only
569 about 100 students per school and SAT scores explaining over 80% of the
570 variance in the school level means on our measures.

571 The modest correlations among individual open-ended tasks that we
572 and others have found (e.g., see Erwin and Schrell, 2003) could pose a
573 problem if the goal was to treat these tasks as parallel forms of the same
574 test or to use the scores on a single task to make important decisions about
575 individual students. However, score reliability was more than adequate for
576 the purposes of reporting results for colleges. Moreover, by using the
577 matrix sampling approach utilized in this study, institutions can measure
578 critical thinking and writing skills across a much broader spectrum of
579 academic disciplines than would be feasible with a single task.

580 Finally, our cross-sectional analyses found that student performance on
581 our measures was related to grades in college and after controlling on SAT
582 scores, mean scores increased consistently from freshman to senior class.
583 These findings suggest that our measures are sensitive to student learning
584 over time. However, as is true of any large-scale assessment (e.g., NAEP),
585 we cannot say whether this improvement was due to college experience,
586 some other experience, maturation, or some combination of these or other
587 factors, although the first attribution seems most plausible. Similarly,
588 unmeasured variables (such as differences in student motivation across
589 colleges) may have contributed to the school effects we observed (see
590 McCaffrey et al., 2003 for a discussion of the limitations of value added
591 analyses). Nevertheless, given the increasing pressure to hold colleges
592 accountable, it would be better to evaluate colleges on the basis of an
593 imperfect measure of student learning than on indicators that have little or
594 no relationship with grades.

595 One of the major challenges to scaling up the approach used in this
596 study is finding effective ways to motivate students to participate. We paid
597 the students in this research, but that is not feasible for a large-scale
598 assessment program. It also raises concern about possible sample selection
599 bias. An operational program would therefore have to embed the
600 measures in capstone courses or all students would have to take them to
601 satisfy graduation requirements. In the focus groups that were held after
602 the test sessions, students reported that the intrinsically interesting nature
603 of the tasks encouraged them to try their best, so motivating them to do
604 well may be less of a challenge than getting them to take the measures.

605 IMPLICATIONS

606 This study describes the results with a promising set of cognitive
607 assessment tools that can be used to measure general as distinct from

608 domain-specific reasoning and writing abilities that are valued as
609 outcomes of college attendance. These open-ended measures can be
610 administered in a few hours and scored reliably. A machine can even grade
611 some and perhaps eventually all of the answers. This makes these
612 measures very efficient relative to other outcomes assessment batteries.
613 The measures appear to assess important abilities that are applicable
614 across major fields. In addition, the scores on them appear to be sensitive
615 to between-institution effects. Such findings are rare in the higher
616 education research literature (Pascarella and Terenzini, 1991). These
617 instruments may also prove useful for benchmarking purposes if future
618 studies with larger numbers of institutions replicate our findings of
619 statistically significant between-institution effects.

620 One of our next steps will be to investigate how scores on the outcome
621 measures used in this research relate to engagement and participation
622 measures at both the student and school levels. This will allow us to
623 estimate the extent to which those practices that the literature espouses
624 to be educationally effective (Chickering and Gamson, 1987; Kuh, 2001,
625 2003) translate into cognitive payoffs. With data from enough colleges
626 and universities it also may be possible to develop residual models,
627 whereby we can compare how particular institutions or types of
628 colleges actually score with how they would be predicted to score, given
629 the nature of their students and institutional characteristics. This would
630 open up a potentially instructive approach to measuring institutional
631 effectiveness. In addition, combining these value-added measures with
632 other information about students (e.g., NSSE) and institutions will allow
633 us to learn much more about the impact of college on student learning as
634 well as the kinds of educational experiences that contribute to desired
635 college outcomes.

636 Finally, we found that the measures we used produced reasonably
637 reliable student-level scores that correlated as highly with GPAs as did
638 SAT scores. Future studies with larger and more representative samples of
639 institutions and students may therefore find a role for such measures in the
640 college admissions process (see Klein et al., 2004 for a discussion of this
641 topic). Indeed, some of the prompts we used are already an integral part of
642 the GRE and the Graduate Management Admissions Test.

643 ENDNOTES

- 644 1. There are multiple theories of intelligence with Spearman at one extreme postulating a
645 single undifferentiated general intelligence and at the other Guilford postulating 128
646 abilities and Gardner postulating different, independent intelligences. Shavelson and
647 Huang do not intend to resolve this dispute (but see Carroll, 1993 or Gustoffson, 1996 for

- 648 recent treatments). Rather, their intent is heuristic, providing a framework in which to
 649 locate debates and achievement tests that have been used in the past to assess student
 650 learning.
 651 2. See Kuh et al. (2004) for a discussion of the correlation of our measures with various
 653 652 NSSE scales.

654 REFERENCES

- 655 Astin, A. W. (1968). Undergraduate achievement and institutional "excellence".
 656 *Science* 161: 661–668.
 657 Astin, A. W. (1977). *Four Critical Years: Effects of College on Beliefs, Values, and*
 658 *Knowledge*. San Francisco: Jossey-Bass.
 659 Astin, A. W. (1991). *Assessment for Excellence: The Philosophy and Practice of*
 660 *Assessment and Evaluation in Higher Education*. New York: American Council on
 661 Education/Macmillan.
 662 Astin, A. W. (1993). *What Matters in College? Four Critical Years Revisited*. San
 663 Francisco: Jossey-Bass.
 664 Banta, T. W., Lund, J. P., and Oblander, F. W. (eds.) (1996). *Assessment in Practice:*
 665 *Putting Principles to Work on College Campuses*. San Francisco: Jossey-Bass.
 666 Benjamin, R., and Hersh, R. H. (2002). Measuring the difference college makes: The
 667 RAND/CAE value added assessment initiative. *Peer Review* 4: 7–10.
 668 Black, S. (1993). Portfolio Assessment. *The Executive Educator* 15: 28–31.
 669 Bohr, L., Pascarella, E., Nora, A., Zusman, B., Jacobs, M., Desler, M., and
 670 Bulakowski, C. (1994). Cognitive effects of two-year and four-year institutions:
 671 a preliminary study. *Community College Review* 22(1): 411.
 672 Bransford, J. D., Brown, A. L., and Cocking, L. L. (1999). *How People Learn: Brain,*
 673 *Mind, Experience, and School*. Washington, DC: National Academy Press.
 674 Burke, J. C., and Minassians, H. (2002). *Performance Reporting: The Preferred "No*
 675 *Cost" Accountability Program* (2001). Albany: The Nelson A. Rockefeller Institute
 676 of Government.
 677 Callan, P. M., and Finney, J. E. (2002). Assessing educational capital: an imperative for
 678 policy. *Change* (34): 25–31.
 679 Carroll, J. B. (1993). Human Cognitive Abilities. *A Survey of Factor-Analytic Studies*.
 680 Cambridge, England: Cambridge University Press.
 681 Chickering, A. W., and Gamson, Z. F. (1987). Seven principles for good practice in
 682 undergraduate education. *American Association for Higher Education Bulletin* 39(7):
 683 3–7.
 684 Cole, J. J. K., Nettles, M. T., and Sharp, S. (1997). *Assessment of teaching and learning*
 685 *for improvement and accountability: state governing, coordinating board and regional*
 686 *accreditation association policies and practices*. Ann Arbor: University of Michigan,
 687 National Center for Postsecondary Improvement.
 688 Cronbach, L. J. (ed.) (2000). *Remaking the Concept of Aptitude: Extending the Legacy*
 689 *of Richard E. Snow*. Mahway, NJ: Erlbaum.
 690 Dey, E., Hurtado, S., Rhee, B., Inkelas, K. K., Wimsatt, L. A., and Guan, F. (1997).
 691 *Improving Research on Postsecondary Outcomes: A Review of the Strengths and*
 692 *Limitations of National Data Sources*. Stanford, CA: National Center for Post-
 693 secondary Improvement.
 694 Erwin, T. D., and Schrell, K. W. (2003). Assessment of critical thinking: New Jersey's
 695 tasks in critical thinking. *The Journal of General Education* 52: 50–70.

- 696 Ewell, P. T. (1984). *The Self-regarding Institution: Information for Excellence*. Boulder,
697 CO: National Center for Higher Education Management Systems.
- 698 Ewell, P. T. (1987). Establishing a campus-based assessment program. In Halpern, D.
699 F. (ed.), *Student outcomes assessment: what institutions stand to gain*. *New Direc-*
700 *tions for Higher Education* 59: 9–24.
- 701 Ewell, P. T. (1988). Outcomes, assessment, and academic improvement: In search of
702 usable knowledge. In Smart, J. C. (ed.), *Higher Education: Handbook of Theory and*
703 *Research*, Vol. IV, pp. 53–108. New York: Agathon Press.
- 704 Ewell, P. T. (1994). *A Policy Guide for Assessment: Making Good Use of the Tasks in*
705 *Critical Thinking*. Princeton, NJ: Educational Testing Service.
- 706 Flowers, L., Osterlind, S. J., Pascarella, E. T., and Pierson, C. T. (2001). How much do
707 students learn in colleges? *The Journal of Higher Education* 72(5): 565–583.
- 708 Fong, B. (1988). Assessing the departmental major. In McMillan, J. H. (ed.), *Assessing*
709 *Students' Learning. New Directions for Teaching and Learning*, Vol. 34, pp. 71–83.
710 San Francisco: Jossey-Bass.
- 711 Forrest, A. (1990). *Time Will Tell: Portfolio-assisted Assessment of General Education*.
712 The AAHE Assessment Forum, American Association for Higher Education.
- 713 Gates, S. M., Augustine, C., Benjamin, R., Bikson, T., Derghazarian, E., Kaganoff, T.,
714 Levy, D., Moini, J., and Zimmer, R. (2001). *Ensuring the quality and productivity of*
715 *education and professional development activities: a review of approaches and lessons*
716 *for DoD*. Santa Monica, CA: National Defense Research Institute, RAND.
- 717 Gentemann, K. M., Fletcher, J. J., and Potter, D. L. (1994). Refocusing the academic
718 program review on student learning. In Kinnick, M. K. (ed.), *Providing useful in-*
719 *formation for deans and department chairs*, *New Directions for Institutional Re-*
720 *search*, No. 84, pp. 31–46. Jossey-Bass: San Francisco.
- 721 Gill, W. E. (1993). *Conversations about accreditation: Middle States Association of*
722 *Colleges and Schools: Focusing on outcomes assessment in the accreditation process*,
723 Paper presented at Double Feature Conference on Assessment and Continuous
724 Quality Improvement of the American Association for Higher Education. Chicago,
725 IL. (ERIC Document Reproduction Service No. ED 358 792).
- 726 Graham, A., and Thompson, N. (2001). *Broken ranks: U.S. News' college rankings*
727 *measure everything but what matters. And most universities do not seem to mind*. The
728 *Washington monthly*. Available at: [www.washingtonmonthly.com/features/2001/](http://www.washingtonmonthly.com/features/2001/0109.graham.thompson.html)
729 [0109.graham.thompson.html](http://www.washingtonmonthly.com/features/2001/0109.graham.thompson.html).
- 730 Gustafsson, J. E., and Undheim, J. O. (1996). Individual differences in cognitive
731 functions. In Calfee, R., and Berliner, D. (eds.), *Handbook of Educational Psychol-*
732 *ogy*. New York: Macmillan, pp. 186–242.
- 733 Halpern, D. F. (1987). Recommendations and caveats. In Halpern, D. F. (ed.), *Student*
734 *Outcomes Assessment: What Institutions Stand to Gain. New directions for higher*
735 *education*, Vol. 59, pp. 109–111. San Francisco: Jossey-Bass.
- 736 Hutchings, P. (1989). *Behind outcomes: contexts and questions*. *The AAHE Assessment*
737 *Forum*, American Association for Higher Education.
- 738 Immerwahl, J. (2000). *Great Expectations: How Californians View Higher Education*.
739 National Center for Public Policy and Higher Education and Public Agenda, San
740 Jose, CA (Table 3, National Column).
- 741 Jacobi, M., Astin, A., and Ayala, F. (1987). *College student outcomes assessment: a*
742 *talent development perspective*. Association for the Study of Higher Education,
743 Washington, DC (ASHE-ERIC Higher Education Report No. 7).
- 744 Johnson, R., McCormick, R. D., Prus, J. S., and Rogers, J. S. (1993). Assessment
745 options for the college major. In Banta, T. W., and Associates (eds.), *Making a*

- 746 *Difference: Outcomes of a Decade of Assessment in Higher Education*, pp. 151–167.
 747 San Francisco: Jossey-Bass.
- 748 Klein, S. (1996). The costs and benefits of performance testing on the bar examination.
 749 *The Bar Examiner* 65(3): 13–20.
- 750 Klein, S., and Hamilton, L. (1998). *The validity of the U.S. News and World Report*
 751 *ranking of ABA law schools*, Report commissioned by the Association of American
 752 Law Schools (available on the web at <http://www.aals.org/validity.html>).
- 753 Klein, S. (2001). *Rationale and plan for assessing higher education outcomes with direct*
 754 *constructed response measures of student skills*. New York, NY: Council for Aid to
 755 Education, Higher Education Policy Series, Number 3.
- 756 Klein, S. (2002). Direct assessment of cumulative student learning. *Peer Review* 4: 26–28.
- 757 Klein, S., Kuh, G., Chun, M., Hamilton, L., and Shavelson, R. (2003). The search for
 758 “Value-Added”: Assessing and validating selected higher education outcomes. Paper
 759 presented at the meetings of the American Educational Research Association,
 760 Chicago, Illinois.
- 761 Klein, S., Shavelson, R., Hamilton, L., and Chun, M. (2004). *Characteristics of hand*
 762 *and machine-assigned scores to college students’ answers to open-ended tasks* (un-
 763 published report).
- 764 Kyllonen, P.C., and Shute, V.J. (1989). A taxonomy of learning skills. In Ackerman,
 765 P. L., Sternberg, R. J., and Glaser, R. (eds.), *Learning and Individual Differences:*
 766 *Advances in Theory and Research*, pp. 117–163. New York: Freeman.
- 767 Kuh, G. D. (2001). Assessing what really matters to student learning: inside the Na-
 768 tional Survey of Student Engagement. *Change* 33(3): 10–17, 66.
- 769 Kuh, G. D. (2003). What we’re learning about student engagement from NSSE.
 770 *Change* 35(2): 24–32.
- 771 Kuh, G., Carini, R., and Klein, S. (2004). *Student engagement and student learning:*
 772 *insights from a construct validation study*. San Diego, California: Paper presented at
 773 the meetings of the American Educational Research Association.
- 774 Lenning, O. T. (1988). Use of noncognitive measures in assessment. In Banta, T. W.
 775 (ed.), *Implementing Outcomes Assessment: Promise and Perils. New Directions for*
 776 *Institutional Research*, Vol. 59, pp. 41–51. San Francisco: Jossey-Bass.
- 777 Machung, A. (1995). *Changes in college rankings: How real are they?* Paper presented at
 778 the 35th Annual AIR Forum, Boston, MA.
- 779 Martinez, M. E. (2000). *Education as the Cultivation of Intelligence*. Mahway, NJ:
 780 Erlbaum.
- 781 McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., and Hamilton, L. S. (2003). *Evalu-*
 782 *ating Value-added Models for Teacher Accountability*. Santa Monica, CA: RAND.
- 783 McGuire, M. D. (1995). Validity issues for reputational studies. In Walleri, R. D., and
 784 Moss, M. K. (eds.), *Evaluating and Responding to College Guidebooks and Rankings.*
 785 *New Directions for Institutional Research*, Vol. 88. San Francisco: Jossey-Bass.
- 786 Messick, S. (1984). The Psychology of Educational Measurement. *Journal of Educa-*
 787 *tional Measurement* 21(3): 215–237.
- 788 Muffo, J. A., and Bunda, M. A. (1993). Attitude and Opinion Data. In Banta, T., and
 789 Associates (eds.), *Making a difference: Outcomes of a decade of assessment in higher*
 790 *education*, pp. 139–150. San Francisco: Jossey-Bass.
- 791 National Center for Higher Education Management Systems (NCEMS) (1994). *A*
 792 *preliminary study of the feasibility and utility for national policy of instructional and*
 793 *good practice indicators in undergraduate education*. Contractor Report for the
 794 National Center for Education Statistics. Boulder, CO: National Center for Higher
 795 Education Management Systems.

- 796 National Center for Higher Education Management Systems (NCEMS) (1996). *The*
797 *National Assessment of College Student Learning: An Inventory of State-level Assessment*
798 *Activities*, Boulder, CO: National Center for Higher Education Management Systems.
- 799 National Opinion Research Center (1997). *A review of the methodology for the U.S.*
800 *News and World Report's rankings of undergraduate colleges and universities*. Report
801 by the National Opinion Research Center.
- 802 National Postsecondary Education Cooperative (2000a). *The NPEC sourcebook on*
803 *assessment, volume 1: Definitions and assessment methods for critical thinking, pro-*
804 *blem solving, and writing*. Center for Assessment and Research Studies, James
805 Madison University, Harrisonburg, VA, under the sponsorship of the National
806 Center for Education Statistics, U.S. Department of Education.
- 807 National Postsecondary Education Cooperative (2000b). *The NPEC sourcebook on*
808 *assessment, volume 2: Selected institutions utilizing assessment results*. Center for
809 Assessment and Research Studies, James Madison University, Harrisonburg, VA,
810 under the sponsorship of the National Center for Education Statistics, U.S.
811 Department of Education.
- 812 Naughton, B. A., Suen, A. Y., and Shavelson, R. J. (2003). *Accountability for what?*
813 *Understanding the learning objectives in state higher education accountability*
814 *programs*. Paper presented at the annual meetings of the American Educational
815 Research Association, Chicago.
- 816 Obler, S. S., Slark, J., and Umbdenstock, L. (1993). Classroom assessment. In Banta,
817 T. W., and Associates (eds.), *Making a difference: Outcomes of a decade of assess-*
818 *ment in higher education*, pp. 211–226. San Francisco: Jossey-Bass.
- 819 Pace, C. R. (1990). *The undergraduates: A report of their activities and progress in*
820 *college in the 1980's*. Los Angeles: Center for the Study of Evaluation, University of
821 California, Los Angeles.
- 822 Palomba, C. A., and Banta, T. W. (1999). *Assessment essentials: Planning, im-*
823 *plementing, and improving assessment in higher education*. San Francisco: Jossey-
824 Bass.
- 825 Pascarella, E. T., and Terenzini, P. T. (1991). *How college affects students: Findings and*
826 *insights from twenty years of research*. San Francisco: Jossey-Bass.
- 827 Pascarella, E. T., Bohr, L., Nora, A., and Terenzini, P.T. (1996). “Is differential
828 exposure to college linked to the development of critical thinking?”. *Research in*
829 *Higher Education* 37: 159–174.
- 830 Pascarella, E. T. (2001). Cognitive growth in college. *Change* 33: 21–27.
- 831 Pellegrino, J. W., Chudowsky, N., and Glaser, R. eds. (2001). *Knowing What Students*
832 *Know: The Science and Design of Educational Assessment*. Washington, DC:
833 National Academy Press.
- 834 Ratcliff, J. L., Jones, E. A., Guthrie, D. S., and Oehler, D. (1991). *The effect of*
835 *coursework patterns, advisement, and course selection on the development of*
836 *general learned abilities of college graduates*. University Park: The Pennsylvania
837 State University, National Center on Postsecondary Teaching, Learning, and
838 Assessment.
- 839 Ratcliff, J. L., and Jones, E. A. et al. (1997). *Turning results into improvement strategies*.
840 The Pennsylvania State University, National Center on Postsecondary Teaching,
841 Learning, and Assessment, University Park.
- 842 Riggs, M. L., and Worthley, J. S. (1992). Baseline Characteristics of Successful
843 Program of Student Outcomes Assessment, ERIC document ED353285.
- 844 Shavelson, R.J., Roeser, R.W., Kupermintz, H., Lau, S., Ayala, C., Haydel, A.,
845 Schultz, S., Quihuis, G., and Gallagher, L. (2002). Richard E. Snow's remaking of

- 846 the concept of aptitude and multidimensional test validity: introduction to the
847 special issue. *Educational Assessment* 8(2): 77–100.
- 848 Shavelson, R.J., and Huang, L. (2003). Responding responsibly to the frenzy to assess
849 learning in higher education. *Change* 35(1): 10–19.
- 850 Smith, M. K., Bradley, J. L., and Draper, G. F. (1993). *A National Survey on*
851 *Assessment Practices*. Knoxville, TN: University of Tennessee, Knoxville, Clear-
852 ingtonhouse for Higher Education Assessment Instruments.
- 853 Snow, R. E. (1994). Abilities in Academic Tasks. In Sternberg, R. J., and Wagner, R.
854 K. (eds.), *Mind in Context: Interactionist Perspectives on Human Intelligence*.
855 Cambridge, England: Cambridge University Press, p. 337.
- 856 Snow, R. E., and Lohman, D. F. (1989). Implications of cognitive psychology for
857 educational measurement. In Linn, R. (ed.), *Educational Measurement*, 3rd ed.,
858 pp. 263–331. New York: Macmillan.
- 859 Steele, J. M., and Lutz, D. A. (1995). *Report of ACT's research on postsecondary*
860 *assessment needs*. Iowa City, IA: American College Testing Program.
- 861 Suen, H. K., and Parkes, J. (1996). Challenges and opportunities for student assessment
862 in distance education. *Distance Education Online Symposium* 6(7): [On-line serial].
863 Available: Internet: ACSDE@PSUVM.PSU.EDU.
- 864 Terenzini, P. T., and Wright, T. (1987). Influences on students' academic growth during
865 four years of college. *Research in Higher Education* 26: 161–179.
- 866 Terenzini, P. T. (1989). Assessment with open eyes: pitfalls in studying student out-
867 comes. *Journal of Higher Education* 60: 644–664.
- 868 Vandament, W. E. (1987). A state university perspective on student outcomes assess-
869 ment. In Halpern, D. F. (ed.), *Student outcomes assessment: what institutions stand*
870 *to gain*. *New Directions for Higher Education*, 59: 25–28
- 871 Waluconis, C. J. (1993). Student self-evaluation. In Trudy, B. (ed.), *Making a differ-*
872 *ence: Outcomes of a decade of assessment in higher education*, pp. 244–255. San
873 Francisco: Jossey-Bass.
- 874 Winter, D. G., McClelland, D. C., and Stewart, A. J. (1981). *A new case for the liberal*
875 *arts*. San Francisco: Jossey-Bass.
- 876 Received November 10, 2003.