Research in Higher Education, Vol. 46, No. 3, May 2005 (© 2005)
DOI: 10.1007/s11162-004-1640-3

AN APPROACH TO MEASURING COGNITIVE OUTCOMES ACROSS HIGHER EDUCATION INSTITUTIONS

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

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

KEY WORDS: value added; assessment; measuring student outcomes.

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

*The RAND Corporation, Santa Monica, California.
**Indiana University Center for Postsecondary Research, Bloomington, IN.
***Council for Aid to Education, New York City, NY.
†The RAND Corporation, Pittsburgh, PA.
‡Graduate School of Education and Department of Psychology (by courtesy), 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

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

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

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

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

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

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

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

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

BACKGROUND: PAST EFFORTS TO MEASURE STUDENT LEARNING

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

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

Actuarial Data

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

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

Ratings of Institutional Quality

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

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

Student Surveys

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

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

Direct Assessments of Student Learning

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

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

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

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

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

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

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

CONCEPTUAL FRAMEWORK

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

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

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

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

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

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

Shavelson and Huang acknowledged that their hierarchy oversimplifies. Knowledge and abilities are interdependent. Learning depends not only on instruction but also on the knowledge and abilities students bring to college. Indeed, instruction and abilities are likely to interact to produce learning, and the course of this interaction evolves over time so that different abilities are called forth and different learning tasks are needed in 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 *across* broad domains (humanities, social sciences, sciences)	GRE writing prompts: make and break an argument
V	Reasoning, comprehending, problem solving, and decision making *within* 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.

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

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

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

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

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

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

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

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

INVESTIGATING THE FEASIBILITY OF USING DIRECT MEASURES

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

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

Participating Institutions

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

Sampling Students Within Colleges

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

Measures

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

Graduate Record Examination (GRE) Essay Prompts

The GRE now includes two essay questions. The 45-minute ''make-an-argument'' 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 (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

[a]Public funding also indicates state controlled.

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

Task Evaluation Form

This questionnaire asked students about the appropriateness and other characteristics of the tests 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 or ACT scores, academic major, college Grade Point Average (GPA), years attending the school, and credit hours earned.

National Survey of Student Engagement (NSSE)

The NSSE has four parts. One part asks students about 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 students' 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 620,000 students at 850 different four-year colleges and universities have completed the NSSE survey since 2000.

Research Design and Test Administration

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

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

		GRE Tasks	
Set	90-Minute Task	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.

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

All of the tests were administered in a controlled setting, usually in one of the college'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 3–3.5 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.

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

Scaling

To combine results across colleges, 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 college) to adjust the correlations with GPAs for possible differences in grading standards among colleges. 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.

Scoring

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

Score Reliability

Table 5 shows that there was a very high degree of consistency 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; the mean correlation between any two of them was .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 an individual student's total score for a 3-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. The reliability of a college's mean score would exceed .90.

Response Mode Effects

A regression analysis that predicted a student's score on our 90-minute 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.

(i.e., handwrite and/or use a computer to draft answers) found that the students who used a computer earned about one third of a standard deviation higher score 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.

Correlations with Other Measures

Student SAT scores and GPAs had somewhat higher correlations with scores on a 3-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 6). Some of this difference can be attributed to the differences in the reliability of the scores from these two batteries.[2]

Hand versus Machine Scoring

At the individual student level, there was a .78 correlation between the hand and machine scoring of the total score across a pair of GRE prompts. 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.

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

Changes in Performance from Freshman to Senior Year

To explore whether our open-ended measures were sensitive to changes in student ability over time (i.e., from freshman to senior year), 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 tasks that the student took. This analysis (which controlled for the student's SAT score, school, and gender) found that the average scores on our measures increased with each class. Specifically, 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 11 colleges where testing was done in the spring of 2002.

School Effects

To study the potential impact of college on student achievement, we regressed the mean score for each college on its average SAT score. This equation also had a dummy variable for each college. This analysis found that mean SAT scores by themselves explained about 82% of the variance in mean college achievement scores. Despite this strong correlation and the modest sample sizes, three colleges 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. Figure 1 shows this relationship (there is one data point plotted for each school).

TABLE 7. Different GRE Scoring Methods Yield Similar Correlations

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

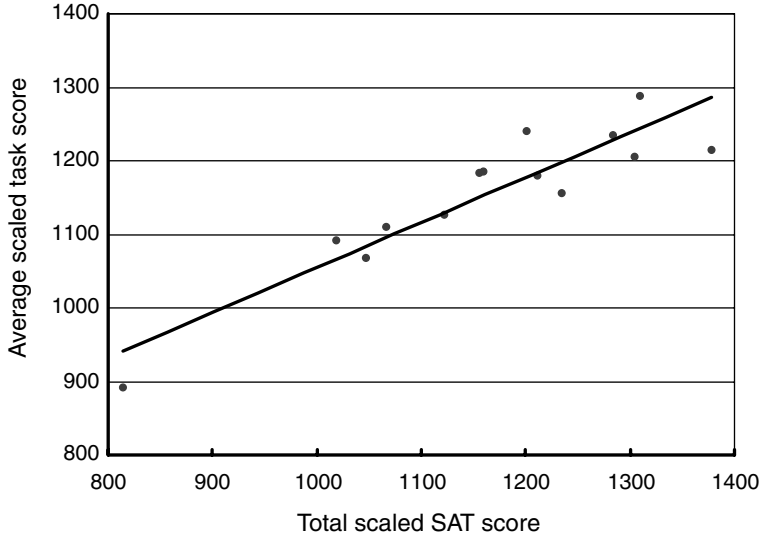

FIG. 1. Relationship between a College's mean SAT and scaled task score.

Student Evaluations of Tasks

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

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

CONCLUSIONS

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

We found that student answers on our measures can be scored very 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?"

		90-Minute Critical Thinking Performance Tasks					
Rating Scale	GRE Writing	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?"

		90-Minute Critical Thinking Performance Tasks					
Rating Scale	GRE Writing	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

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

Our results further suggest that a 3-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% of the variance in the school level means on our measures.

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

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

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

IMPLICATIONS

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

domain-specific reasoning and writing abilities that are valued as outcomes of 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 abilities 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 and 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 the outcome measures used in this research relate to engagement and participation 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 and 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., 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 the measures we used produced reasonably reliable student-level scores that correlated as highly with GPAs as did SAT scores. Future studies with larger and more representative samples of institutions and students may therefore find a role for such measures in the college admissions process (see Klein et al., 2004 for a discussion of this topic). Indeed, some of the prompts we used are already an integral part of the GRE and the Graduate Management Admissions Test.

ENDNOTES

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

recent treatments). Rather, their intent is heuristic, providing a framework in which to locate debates and achievement tests that have been used in the past to assess student learning.

2. See Carini et al. (2004) for a discussion of the correlation of our measures with various NSSE scales.

REFERENCES

Astin, A. W. (1968). Undergraduate achievement and institutional "excellence". *Science* 161: 661–668.

Astin, A. W. (1977). *Four Critical Years: Effects of College on Beliefs, Values, and Knowledge*. San Francisco: Jossey-Bass.

Astin, A. W. (1991). *Assessment for Excellence: The Philosophy and Practice of Assessment and Evaluation in Higher Education*. New York: American Council on Education/Macmillan.

Astin, A. W. (1993). *What Matters in College? Four Critical Years Revisited*. San Francisco: Jossey-Bass.

Banta, T. W., Lund, J. P., and Oblander, F. W. (eds.) (1996). *Assessment in Practice: Putting Principles to Work on College Campuses*. San Francisco: Jossey-Bass.

Benjamin, R., and Hersh, R. H. (2002). Measuring the difference college makes: The RAND/CAE value added assessment initiative. *Peer Review* 4: 7–10.

Black, S. (1993). Portfolio Assessment. *The Executive Educator* 15: 28–31.

Bohr, L., Pascarella, E., Nora, A., Zusman, B., Jacobs, M., Desler, M., and Bulakowski, C. (1994). Cognitive effects of two-year and four-year institutions: a preliminary study. *Community College Review* 22(1): 411.

Bransford, J. D., Brown, A. L., and Cocking, L. L. (1999). *How People Learn: Brain, Mind, Experience, and School*. Washington, DC: National Academy Press.

Burke, J. C., and Minassians, H. (2002). *Performance Reporting: The Preferred "No Cost" Accountability Program* (2001). Albany: The Nelson A. Rockefeller Institute of Government.

Callan, P. M., and Finney, J. E. (2002). Assessing educational capital: an imperative for policy. *Change* (34): 25–31.

Carini, R., Kuh, G., and Klein, S. (2004). *Student engagement and student learning: insights from a construct validation study*. San Diego, California: Paper presented at the meetings of the American Educational Research Association.

Carroll, J. B. (1993). Human Cognitive Abilities. *A Survey of Factor-Analytic Studies*. Cambridge, England: Cambridge University Press.

Chickering, A. W., and Gamson, Z. F. (1987). Seven principles for good practice in undergraduate education. *American Association for Higher Education Bulletin* 39(7): 3–7.

Cole, J. J. K., Nettles, M. T., and Sharp, S. (1997). *Assessment of teaching and learning for improvement and accountability: state governing, coordinating board and regional accreditation association policies and practices*. Ann Arbor: University of Michigan, National Center for Postsecondary Improvement.

Cronbach, L. J. (ed.) (2000). *Remaking the Concept of Aptitude: Extending the Legacy of Richard E. Snow*. Mahway, NJ: Erlbaum.

Dey, E., Hurtado, S., Rhee, B., Inkelas, K. K., Wimsatt, L. A., and Guan, F. (1997). *Improving Research on Postsecondary Outcomes:A Review of the Strengths and*

Limitations of National Data Sources. Stanford, CA: National Center for Postsecondary Improvement.
Erwin, T. D., and Schrell, K. W. (2003). Assessment of critical thinking: New Jersey's tasks in critical thinking. *The Journal of General Education* 52: 50–70.
Ewell, P. T. (1984). *The Self-regarding Institution: Information for Excellence*. Boulder, CO: National Center for Higher Education Management Systems.
Ewell, P. T. (1987). Establishing a campus-based assessment program. In Halpern, D. F. (ed.), Student outcomes assessment: what institutions stand to gain. *New Directions for Higher Education* 59: 9–24.
Ewell, P. T. (1988). Outcomes, assessment, and academic improvement: In search of usable knowledge. In Smart, J. C. (ed.), *Higher Education: Handbook of Theory and Research*, Vol. IV, pp. 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.
Flowers, L., Osterlind, S. J., Pascarella, E. T., and Pierson, C. T. (2001). How much do students learn in colleges? *The Journal of Higher Education* 72(5): 565–583.
Fong, B. (1988). Assessing the departmental major. In McMillan, J. H. (ed.), *Assessing Students' Learning*. *New Directions for Teaching and Learning*, Vol. 34, pp. 71–83. San Francisco: Jossey-Bass.
Forrest, A. (1990). *Time Will Tell: Portfolio-assisted Assessment of General Education*. The AAHE Assessment Forum, American Association for Higher Education.
Gates, S. M., Augustine, C., Benjamin, R., Bikson, T., Derghazarian, E., Kaganoff, T., Levy, D., Moini, J., and Zimmer, R. (2001). *Ensuring the quality and productivity of education and professional development activities: a review of approaches and lessons for DoD*. Santa Monica, CA: National Defense Research Institute, RAND.
Gentemann, K. M., Fletcher, J. J., and Potter, D. L. (1994). Refocusing the academic program review on student learning. In Kinnick, M. K. (ed.), Providing useful information for deans and department chairs, New Directions for Institutional Research, No. 84, pp. 31–46. Jossey-Bass: San Francisco.
Gill, W. E. (1993). *Conversations about accreditation: Middle States Association of Colleges and Schools: Focusing on outcomes assessment in the accreditation process*, Paper presented at Double Feature Conference on Assessment and Continuous Quality Improvement of the American Association for Higher Education. Chicago, IL. (ERIC Document Reproduction Service No. ED 358 792).
Graham, A., and Thompson, N. (2001). *Broken ranks: U.S. News' college rankings measure everything but what matters. And most universities do not seem to mind*. The Washington monthly. Available at: www.washingtonmonthly.com/features/2001/0109.graham.thompson.html.
Gustafsson, J. E., and Undheim, J. O. (1996). Individual differences in cognitive functions. In Calfee, R., and Berliner, D. (eds.), *Handbook of Educational Psychology*. New York: Macmillan, pp. 186–242.
Halpern, D. F. (1987). Recommendations and caveats. In Halpern, D. F. (ed.), *Student Outcomes Assessment: What Institutions Stand to Gain. New directions for higher education*, Vol. 59, pp. 109–111. San Francisco: Jossey-Bass.
Hutchings, P. (1989). *Behind outcomes: contexts and questions. The AAHE Assessment Forum*, American Association for Higher Education.
Immerwahl, J. (2000). *Great Expectations: How Californians View Higher Education*. National Center for Public Policy and Higher Education and Public Agenda, San Jose, CA (Table 3, National Column).

Jacobi, M., Astin, A., and Ayala, F. (1987). *College student outcomes assessment: a talent development perspective*. Association for the Study of Higher Education, Washington, DC (ASHE-ERIC Higher Education Report No. 7).

Johnson, R., McCormick, R. D., Prus, J. S., and Rogers, J. S. (1993). Assessment options for the college major. In Banta, T. W., and Associates (eds.), *Making a Difference: Outcomes of a Decade of Assessment in Higher Education*, pp. 151–167. San Francisco: Jossey-Bass.

Klein, S. (1996). The costs and benefits of performance testing on the bar examination. *The Bar Examiner* 65(3): 13–20.

Klein, S., and Hamilton, L. (1998). *The validity of the U.S. News and World Report ranking of ABA law schools*, Report commissioned by the Association of American Law Schools (available on the web at http://www.aals.org/validity.html).

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.

Klein, S., Kuh, G., Chun, M., Hamilton, L., and Shavelson, R. (2003). The search for "Value-Added": Assessing and validating selected higher education outcomes. Paper presented at the meetings of the American Educational Research Association, Chicago, Illinois.

Klein, S., Shavelson, R., Hamilton, L., and Chun, M. (2004). *Characteristics of hand and machine-assigned scores to college students' answers to open-ended tasks* (unpublished report).

Kyllonen, P.C., and Shute, V.J. (1989). A taxonomy of learning skills. In Ackerman, P. L., Sternberg, R. J., and Glaser, R. (eds.), *Learning and Individual Differences: Advances in Theory and Research*, pp. 117–163. New York: Freeman.

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.

Lenning, O. T. (1988). Use of noncognitive measures in assessment. In Banta, T. W. (ed.), *Implementing Outcomes Assessment: Promise and Perils. New Directions for Institutional Research*, Vol. 59, pp. 41–51. San Francisco: Jossey-Bass.

Machung, A. (1995). *Changes in college rankings: How real are they?* Paper presented at the 35th Annual AIR Forum, Boston, MA.

Martinez, M. E. (2000). *Education as the Cultivation of Intelligence*. Mahway, NJ: Erlbaum.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., and Hamilton, L. S. (2003). *Evaluating Value-added Models for Teacher Accountability*. Santa Monica, CA: RAND.

McGuire, M. D. (1995). Validity issues for reputational studies. In Walleri, R. D., and Moss, M. K. (eds.), *Evaluating and Responding to College Guidebooks and Rankings. New Directions for Institutional Research*, Vol. 88. San Francisco: Jossey-Bass.

Messick, S. (1984). The Psychology of Educational Measurement. *Journal of Educational Measurement* 21(3): 215–237.

Muffo, J. A., and Bunda, M. A. (1993). Attitude and Opinion Data. In Banta, T, and Associates (eds.), *Making a difference: Outcomes of a decade of assessment in higher education*, pp. 139–150. San Francisco: Jossey-Bass.

National Center for Higher Education Management Systems (NCEMS) (1994). *A preliminary study of the feasibility and utility for national policy of instructional and good practice indicators in undergraduate education*. Contractor Report for the

National Center for Education Statistics. Boulder, CO: National Center for Higher Education Management Systems.
National Center for Higher Education Management Systems (NCEMS) (1996). *The National Assessment of College Student Learning: An Inventory of State-level Assessment Activities*, Boulder, CO: National Center for Higher Education Management Systems.
National Opinion Research Center (1997). *A review of the methodology for the U.S. News and World Report's rankings of undergraduate colleges and universities*. Report by the National Opinion Research Center.
National Postsecondary Education Cooperative (2000a). *The NPEC sourcebook on assessment, volume 1: Definitions and assessment methods for critical thinking, problem solving, and writing*. Center for Assessment and Research Studies, James Madison University, Harrisonburg, VA, under the sponsorship of the National Center for Education Statistics, U.S. Department of Education.
National Postsecondary Education Cooperative (2000b). *The NPEC sourcebook on assessment, volume 2: Selected institutions utilizing assessment results*. Center for Assessment and Research Studies, James Madison University, Harrisonburg, VA, under the sponsorship of the National Center for Education Statistics, U.S. Department of Education.
Naughton, B. A., Suen, A. Y., and Shavelson, R. J. (2003). *Accountability for what? Understanding the learning objectives in state higher education accountability programs*. Paper presented at the annual meetings of the American Educational Research Association, Chicago.
Obler, S. S., Slark, J., and Umbdenstock, L. (1993). Classroom assessment. In Banta, T. W., and Associates (eds.), *Making a difference: Outcomes of a decade of assessment in higher education*, pp. 211–226. San Francisco: Jossey-Bass.
Pace, C. R. (1990). *The undergraduates; A report of their activities and progress in college in the 1980's*. Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles.
Palomba, C. A., and Banta, T. W. (1999). *Assessment essentials: Planning, implementing, and improving assessment in higher education*. San Francisco: Jossey-Bass.
Pascarella, E. T., and Terenzini, P. T. (1991). *How college affects students: Findings and insights from twenty years of research*. San Francisco: Jossey-Bass.
Pascarella, E. T., Bohr, L., Nora, A., and Terenzini, P.T. (1996). "Is differential exposure to college linked to the development of critical thinking?". *Research in Higher Education* 37: 159–174.
Pascarella, E. T. (2001). Cognitive growth in college. *Change* 33: 21–27.
Pellegrino, J. W., Chudowsky, N., and Glaser, R. eds. (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington, DC: National Academy Press.
Ratcliff, J. L., Jones, E. A., Guthrie, D. S., and Oehler, D. (1991). *The effect of coursework patterns, advisement, and course selection on the development of general learned abilities of college graduates*. University Park: The Pennsylvania State University, National Center on Postsecondary Teaching, Learning, and Assessment.
Ratcliff, J. L., and Jones, E. A. et al. (1997). *Turning results into improvement strategies*. The Pennsylvania State University, National Center on Postsecondary Teaching, Learning, and Assessment, University Park.
Riggs, M. L., and Worthley, J. S. (1992). Baseline Characteristics of Successful Program of Student Outcomes Assessment, ERIC document ED353285.

Shavelson, R.J., Roeser, R.W., Kupermintz, H., Lau, S., Ayala, C., Haydel, A., Schultz, S., Quihuis, G., and Gallagher, L. (2002). Richard E. Snow's remaking of the concept of aptitude and multidimensional test validity: introduction to the special issue. *Educational Assessment* 8(2): 77–100.

Shavelson, R.J., and Huang, L. (2003). Responding responsibly to the frenzy to assess learning in higher education. *Change* 35(1): 10–19.

Smith, M. K., Bradley, J. L., and Draper, G. F. (1993). *A National Survey on Assessment Practices*. Knoxville, TN: University of Tennessee, Knoxville, Clearinghouse for Higher Education Assessment Instruments.

Snow, R. E. (1994). Abilities in Academic Tasks. In Sternberg, R. J., and Wagner, R. K. (eds.), *Mind in Context: Interactionist Perspectives on Human Intelligence*. Cambridge, England: Cambridge University Press, p. 337.

Snow, R. E., and Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In Linn, R. (ed.), *Educational Measurement*, 3rd ed., pp. 263–331. New York: Macmillan.

Steele, J. M., and Lutz, D. A. (1995). *Report of ACT's research on postsecondary assessment needs*. Iowa City, IA: American College Testing Program.

Suen, H. K., and Parkes, J. (1996). Challenges and opportunities for student assessment in distance education. *Distance Education Online Symposium* 6(7): [On-line serial]. Available: Internet: ACSDE@PSUVM.PSU.EDU.

Terenzini, P. T., and Wright, T. (1987). Influences on students' academic growth during four years of college. *Research in Higher Education* 26: 161–179.

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 Halpern, D. F. (ed.), Student outcomes assessment: what institutions stand to gain. *New Directions for Higher Education*, 59: 25–28

Waluconis, C. J. (1993). Student self-evaluation. In Trudy, B. (ed.), *Making a difference: Outcomes of a decade of assessment in higher education*, pp. 244–255. San Francisco: Jossey-Bass.

Winter, D. G., McClelland, D. C., and Stewart, A. J. (1981). *A new case for the liberal arts*. San Francisco: Jossey-Bass.

Received November 10, 2003.