Research in Higher Education, Vol. 37, No. 1, 1996

LIMITATIONS OF USING STUDENTS' SELF-REPORTS OF ACADEMIC DEVELOPMENT AS PROXIES FOR TRADITIONAL ACHIEVEMENT MEASURES

Gary R. Pike

An important issue in national assessment efforts is how best to measure the outcomes of college. While initial discussions about a national collegiate assessment focused on the reliability, validity, and feasibility of using achievement tests to measure student learning, subsequent discussions have raised the possibility of using students' self-reports of academic development as proxies for achievement test scores. The present study examines the stability of the relationships among self-reports and test scores across samples of two- and four-year colleges and universities. Multitrait-multimethod analyses indicated that self-reports and test scores developed from the same set of test specifications do measure the same constructs, although the scores from one type of measurement may not be "substitutable" for scores from the other type of measurement. In addition, the analyses produced ambiguous results concerning the stability of relationships across different types of institutions.

Few dispassionate observers of higher education would argue that American colleges and universities are not facing a crisis in public confidence. The increasing costs of a college education, coupled with reports criticizing the knowledge and skills of college graduates, have raised serious questions about the value of postsecondary education (Ewell, 1991; Pascarella and Terenzini, 1991; Wingspread Group on Higher Education, 1993). In addition, charges of abuse and mismanagement in higher education have undermined the public's faith in the ability of colleges and universities to regulate their own affairs (Ewell, 1994; McClenney, 1993).

Given this crisis of confidence, it is not surprising that external agencies,

Gary R. Pike, Director, Student Life Studies, University of Missouri–Columbia, 3 Parker Hall, Columbia, MO 65211.

Paper presented at the annual meeting of the Association for Institutional Research, Boston, May 29, 1995.

0361-0365/96/0200-0089$09.50/0

including states and accrediting associations, are taking more active roles in requiring that colleges and universities be accountable for their actions and the quality of their education programs (Ewell, 1994; House, 1993). Typical of this activist role is the effort by the federal government to create a national assessment of college students' critical thinking and communication skills (Elliott, 1991).

This paper reports the results of the third in a series of studies designed to evaluate whether self-reports of students' college experiences can serve as proxies for achievement test scores in a national assessment of college student learning. In particular, the present study investigates whether relationships between self-reports and objective measures of achievement "persist" across different types of institutions.

FEDERAL EFFORTS

In 1991, the National Center for Education Statistics (NCES) began hosting a series of study design workshops to examine the feasibility of creating measures of college student achievement similar to the National Assessment of Educational Progress (NAEP). Several participants attending the first workshop voiced reservations about the feasibility of developing a national assessment of college outcomes. Banta (1991), for example, raised questions about whether it would be possible to achieve a consensus about the outcomes that should be assessed, while Dunbar (1991) identified several technical problems with creating an assessment that would provide reliable and valid data for policy decisions. Other participants were more optimistic. Ratcliff (1991) argued that a national assessment was feasible. However, he urged that the development of a national assessment be a long-term project. In the interim, he suggested that alternative measures should be used as proxies for the proposed national assessment.

The National Education Goals Panel Resource Group on Adult Literacy and Lifelong Learning (1991) also recommended that alternatives to a national test be considered seriously. Noting that the development of a national assessment could take as long as five years and cost millions of dollars, the Resource Group argued that self-reports of academic development should be used as proxies for a national assessment and should serve as guides for policy actions.

Participants in the second study design workshop hosted by NCES moved beyond discussing the feasibility of assessing college-level critical thinking and communicating to proposing specific assessment designs and measurement techniques (Daly, 1994; Halpern, 1994; Perkins, Jay, and Tishman, 1994). These proposals ranged from paper-and-pencil measures to computer-administered tests and complex performance-based assessments. Like Ratcliff and the Resource Group, the participants in the second study design workshop recog-

nized that the development of a national assessment of college student learning would be a long and expensive process. These participants again recommended that students' self-reports of their academic development during college be used as proxies for more traditional achievement tests.[1]

PREVIOUS RESEARCH

In a recent report to NCES, the National Center for Higher Education Management Systems (NCHEMS) enumerated four criteria for evaluating the use of self-reports of academic development as proxies for achievement test scores: (1) the measures should represent broad-based outcomes; (2) the measures should represent significant phenomena that can be used to inform policy actions; (3) the measures should reliably covary with other assessments; and (4) the observed relationships should persist across different educational settings (Ewell et al., 1994).

Applying their own criteria to self-report measures, Ewell, Lovell, Dressler, and Jones (1994) found that surveys, such as the College Student Experiences Questionnaire (CSEQ) (Pace, 1987), provided measures of significant, broad-based outcomes that could be used to inform policy actions. Research by Pike (1995), comparing students' responses to CSEQ-like items with scores on the College Basic Academic Subjects Examination (College BASE), provided empirical support for the conclusions of the NCHEMS researchers. However, Pike's research raised questions about whether the covariance between self-reports of academic development during college and scores on the College BASE were sufficiently high to conclude that both sets of items measured the same constructs.

The ambiguity in Pike's (1995) findings was consistent with the results of earlier research. For example, several studies have reported low to moderately high correlations between self-reports of academic development and scores on achievement tests (Anaya, 1992; Astin, 1993; Baird, 1976; Berdie, 1971; Dumont and Troelstrup, 1980; Pohlmann and Beggs, 1974). Berdie (1971), for example, reported correlations ranging from 0.47 to 0.74 for self-reported knowledge and a test about famous people. Similarly, Pohlmann and Beggs (1974) found that correlations between self-reports and tests of course material ranged from 0.52 to 0.67. In contrast, Dumont and Troelstrup (1980) found that correlations between self-reports and scores on the College Outcome Measures Program (COMP) examination were relatively low, ranging from 0.21 to 0.24. Astin (1993) also reported that he found weak to moderate correlations between self-reports and scores on the Graduate Record Examination (GRE) and the National Teacher Examination (NTE).

Pike (1995) advanced two reasons to account for his findings. The first reason, originally suggested by Dumont and Troelstrup (1980), was that generally

there is a poor content overlap between self-reports of student learning and achievement test scores. In their research, Dumont and Troelstrup noted that self-report items tended to measure generic college outcomes, such as effective writing or critical thinking, while the COMP examination tested more specific skills. This same lack of content overlap was found in Astin's self-report questions and items on the GRE and NTE.

Pike noted that a second factor that could influence the magnitude of the correlations between self-reports and test scores is related to differences in the two measurement methods. Astin (1993) noted that standardized achievement tests tend to have high fidelity, but narrow bandwidth. That is, objective tests generally measure achievement very accurately, but over a relatively narrow range of behavior. In contrast, self-reports have lower fidelity, but greater bandwidth. That is, self-reports tend to measure broad arrays of behavior, but they do so at the cost of precision. Pike argued that these measurement differences can give rise to method-specific score variance, thereby attenuating the correlations between self-reports and test scores.

Pike (1995) noted that the relative impact of content overlap and measurement method differences on the relationship between self-reports and test scores is critical. If low correlations between self-reports and test scores are the result of poor content overlap, correlations can be improved by developing sets of measures with higher content correspondence. However, if the low correlations between self-reports and test scores are the result of basic differences in the two measurement methods, creating valid proxies for test scores would be much more difficult.

In a subsequent study, Pike (1994) sought to identify the relative contributions of poor content overlap and measurement method differences to low correlations between self-report and objective measures of student learning. The data for Pike's follow-up study included students' scores on the College Basic Academic Subjects Examination (College BASE) and self-reports of cognitive development derived from the test specifications underlying College BASE. The subjects for this study were 1,587 students from 10 institutions located in the Mid-Atlantic, Southeastern, and Midwestern regions of the United States. Six of the institutions were four-year colleges and universities, while four were community colleges.

Using multitrait-multimethod analyses similar to those in his first study, Pike (1994) found that a common set of content specifications produced substantially higher rates of convergence between self-reports and test scores. He concluded that high content overlap is a key element in developing self-reports that can serve as proxies for test scores. Consistent with his first study, Pike found evidence of measurement differences between self-reports and test scores. However, his follow-up research suggested a very different interpretation of the nature of those measurement differences. Unlike the initial study, which found

that survey and test factors were uncorrelated, the follow-up study found that there was a moderately significant positive correlation between the survey and test factors.

Based on the results of his two studies, Pike was cautiously optimistic that self-reports of learning and development during college could be used as proxies for exiting test scores. He noted, however, that research on the stability of relationships between self-reports and test scores across educational settings has not been established. He concluded that before self-reports can be used as proxies for test scores, research should be conducted to evaluate the convergent and discriminant validity of the two measurement methods across different types of colleges and universities. The goal of the present research is to do just that. Specifically, the present research examines the convergent and discriminant validity of self-reports and test scores across four community colleges and six four-year colleges and universities.

EVALUATION CRITERIA

In his second study, Pike (1994) argued that while the NCHEMS standards represent *necessary* conditions for self-reports to serve as proxies for test scores, they alone are not sufficient. In particular, the third criterion that measures reliably covary is not sufficiently rigorous to establish the validity of using self-reports as proxies for test scores. Scores on two measures of cognitive development may spuriously covary, due to the presence of correlated errors of measurement or because both instruments tap general intellectual abilities, instead of measuring the same educational outcomes. *In order to serve as proxies for achievement test scores, self-reports must measure the same constructs as the achievement tests.* Satisfying this criterion requires evidence of *convergence* (i.e., covariance among different measures of the same educational outcome) and *discrimination* (i.e., a lack of covariance among measures of different educational outcomes) (Cronbach and Meehl, 1955). These two requirements help ensure that observed associations are significant and not the product of either correlated errors of measurement or the undifferentiated measurement of general intellectual ability.

Several different techniques are available for evaluating the convergence and discrimination of two sets of measures (Widaman, 1985). Of these approaches, the analysis of multitrait-multimethod matrices represents an extremely powerful tool (Campbell and Fiske, 1959). An important advantage of reliance on the multitrait-multimethod approach is that it allows a researcher to assess the strength of the *true* relationship between two or more measurement methods, while providing an indication of whether the various methods can differentiate among constructs (Schmitt and Stults, 1986; Widaman, 1985).

Campbell and Fiske (1959) noted that multitrait-multimethod analysis re-

quires that two or more traits (e.g., educational outcome domains) be measured using two or more methods (e.g., self-reports and test scores). Significant correlations among different methods of measuring the same trait provide evidence of convergence, while the absence of significant correlations among different outcomes provides evidence of discrimination. Research using multitrait-multimethod matrices has found that the correlations among different measures of the same trait are usually significant, but moderate, while different traits also tend to be moderately correlated (Fiske, 1982). Thus, the key to evaluating multitrait-multimethod data is the *relative* strength of the relationships representing convergence and discrimination. A more detailed description of the data analyses involved in establishing evidence of convergence and discrimination is provided later in this paper.

In order to satisfy the NCHEMS criterion that relationships "persist" across different types of institutions, multitrait-multimethod analyses must find similar *patterns* of convergence and discrimination across institutions. Pattern invariance is necessary, but not sufficient, to establish that relationships persist across institutions. If the relationships between observed measures and higher-order representations of methods and traits are not the same across different types of institutions, it is likely that different constructs are being measured, even when there is clear evidence of convergence and discrimination (Marsh, 1994). It is also desirable, but not essential, that the relationships among methods and traits be the same across groups of institutions (Byrne, 1989). Here again, a more detailed description of the procedures for evaluating the persistence of relationships across institutions is provided in the discussion of research methods.

RESEARCH METHODS

Subjects

The subjects in this study were 1,568 students from 10 institutions located in the Mid-Atlantic, Southeastern, and Midwestern regions of the United States. Of the total, 740 students (47.2 percent) were from six four-year colleges and universities, and 828 students (52.8 percent) were from four community colleges. Table 1 presents data on gender and the racial/ethnic characteristics of students at the two- and four-year institutions.

An examination of the data in Table 1 reveals slight differences in the percentages of males and females by type of institution. Of the students attending four-year institutions, 47.5 percent were male and 52.5 percent were female. In contrast, 58.2 percent of the students at the community colleges were male and 41.8 percent were female. Although these differences were statistically significant ($\chi^2 = 5.10$; $df = 1$; $p < .05$), the relationship was relatively weak ($\phi = -0.06$).

Approximately 82.3 percent of the students from four-year institutions classi-

TABLE 1. Sex and Racial/Ethnic Characteristics of the Two- and Four-Year College Samples

	Two-Year Colleges	Four-Year Colleges	All Colleges
Sex			
Male	41.8%	47.5%	44.5%
Female	58.2%	52.5%	55.5%
Race/ethnicity			
African-American	13.3%	13.4%	13.3%
Caucasian	81.4%	82.3%	81.8%
Other	5.3%	5.3%	4.9%

fied themselves as Caucasian, 13.4 percent classified themselves as African American, and 4.3 percent classified themselves in some other racial/ethnic category. Similarly, 81.4 percent of the students attending two-year institutions classified themselves as Caucasian, 13.3 percent classified themselves as African American, and 5.3 percent classified themselves in some other racial/ethnic category. No differences in race/ethnicity were found for the two types of institutions.

Instruments

The data used in this study were students' scores on the College Basic Academic Subjects Examination (College BASE) and self-reports of cognitive development derived from the specifications for College BASE. College BASE is a criterion-referenced achievement test focusing on the degree to which students have mastered particular skills and competencies consistent with the completion of general education coursework at a college or university (Osterlind, 1989). The test assesses learning in four subject areas: (1) English, (2) mathematics, (3) science, and (4) social studies. Subject scores are built upon content clusters which, in turn, are based on skills and enabling subskills (Pike, 1992b). For example, English scores are based on two content clusters: (1) reading and literature, and (2) writing. The cluster score for reading and literature is based on skills related to (1) reading analytically, (2) reading critically, and (3) understanding literature (Osterlind, 1989).

Numerical scores are provided for the four subject areas and the nine content clusters in College BASE, while ratings of high, medium, or low are provided for each skill (Osterlind, 1989). The numerical scores have been scaled to have a theoretical mean of 300 and a standard deviation of 65. No numerical scores or ratings are provided for the enabling subskills. Instead, the enabling subskills are used as a guide for the types of items to be included in the test (Osterlind and Merz, 1992). For example, the enabling subskills underlying the skill of

reading critically include (1) ascertaining the meaning of a passage, (2) recognizing the implicit assumptions underlying a passage, and (3) evaluating the ideas presented in a passage to determine their logical validity, their implications, or their relationships to ideas beyond the text (Osterlind, 1989).

The subjects, clusters, and skills assessed by College BASE were derived from the work of the College Board's Educational Equality Project (Osterlind and Merz, 1992). Initial specifications for the test were drawn from the project's report *Academic Preparation for College: What Examinees Need to Know and Be Able to Do* (College Board, 1983). One strength of the skills and competencies outlined in this report is that they provide for relatively broad coverage within particular subject areas (Osterlind and Merz, 1992). An important limitation of these skills and competencies is that they represent college entrance, not exit, abilities. In order to develop appropriate exiting skills, more than 100 faculty representing 50 postsecondary institutions in 20 states helped revise the skills and competencies identified by the Educational Equality Project, modifying them to reflect the general education knowledge and skills expected of college graduates (Osterlind and Merz, 1992).

Research by Pike (1992b) has provided evidence of the construct validity of College BASE as a measure of general education program effectiveness. Pike found that the empirical structure of the test corresponds to the structure outlined in its test specifications. In addition, he found that the test is sensitive to the effects of general education coursework. Most recently, Pike (1995) reported that College BASE scores are related to students' experiences outside the classroom.

In the present study, the nine cluster scores for College BASE were used to represent the subject areas of English, mathematics, science, and social studies. Previous research has reported that the reliability estimates for the cluster scores range from 0.67 for writing to 0.84 for algebra (Pike, 1992a).

The self-reports of cognitive development used in the present research consisted of 28 questions corresponding to enabling subskills on College BASE. Three questions were included for each College BASE content cluster, except fundamental concepts (in science). Four questions were used to represent fundamental science concepts in order to balance the number of physical and biological science items. For each survey question, students were asked to rate themselves in the top 10 percent, above average, average, below average, or in the bottom 10 percent in comparison to other students they knew. Using procedures developed by Armor (1974), factor scores were calculated at the cluster level. Reliability estimates for the factor scores ranged from 0.68 for social science to 0.88 for algebra and for geometry. Factor scores were scaled to have means of 300 and standard deviations of 65.

Table 2 presents reliability estimates, means, standard deviations, and difference (t) tests for the College BASE and self-report measures. An examination

TABLE 2. Reliability Estimates, Means, and *t*-Test Results for the *College BASE* and Self-Report Scales

	Reliability	Two-Year Colleges	Four-Year Colleges	Difference
College BASE Scales				
Reading and Literature	0.76	280.3	294.6	−4.76[c]
		(59.64)	(59.83)	(1.01)
Writing	0.67	285.2	301.3	−5.68[c]
		(56.63)	(55.49)	(1.04)
General Mathematics	0.80	277.4	304.2	−8.38[c]
		(59.05)	(66.74)	(1.28)[c]
Algebra	0.84	300.7	321.7	−6.84[c]
		(59.00)	(62.64)	(1.13)
Geometry	0.75	286.6	315.2	−9.06[c]
		(59.13)	(66.01)	(1.25)[b]
Lab and Field Work	0.78	273.8	308.7	−9.83[c]
		(67.03)	(73.61)	(1.21)[b]
Fundamental Concepts	0.74	280.8	309.0	−8.47[c]
		(68.37)	(63.47)	(1.16)[a]
History	0.77	290.9	307.7	−5.57[c]
		(60.29)	(58.9)	(1.05)
Social Science	0.75	279.4	298.4	−5.88[c]
		(63.66)	(63.65)	(1.00)
Self-Report Scales				
Reading and Literature	0.76	302.8	297.5	1.59
		(63.73)	(66.43)	(1.09)
Writing	0.82	298.3	302.5	−1.27
		(64.30)	(65.56)	(1.04)
General Mathematics	0.71	295.8	304.1	−2.50[a]
		(63.07)	(67.72)	(1.15)[a]
Algebra	0.88	294.5	305.9	−3.46[c]
		(61.37)	(68.48)	(1.25)[b]
Geometry	0.88	293.4	306.9	−4.11[c]
		(61.84)	(67.88)	(1.20)[b]
Lab and Field Work	0.79	290.9	310.4	−6.00[c]
		(64.18)	(64.70)	(1.02)
Fundamental Concepts	0.86	292.6	308.6	−4.90[c]
		(65.17)	(64.02)	(1.04)
History	0.81	300.7	399.3	0.44
		(64.55)	(65.41)	(1.03)
Social Science	0.68	300.2	300.0	0.06
		(65.77)	(64.58)	(1.04)

of the means and *t*-test results reveals that the four-year college means were significantly greater than the corresponding two-year means for every College BASE scale. Four-year college means for the mathematics and science self-report scales also were significantly greater than corresponding two-year means. No significant differences in two- and four-year college means were found for the English and social studies self-report measures.

The standard deviations and tests of homogeneity of variance (shown in parentheses in Table 2) indicated that variances were generally similar across types of institutions. Both the English and social studies College BASE and self-report scales have similar variances for two- and four-year institutions. In contrast, only the mathematics subscales show significant differences across both the College BASE and self-report measures. While there were significant differences in the sample variances for the College BASE science subscales, no significant differences were found for the self-report science scales.

Data Analysis

The data analyses were conducted in two phases. First, separate multitrait-multimethod analyses were conducted for two- and four-year colleges and universities to determine if there was evidence of convergence and discrimination within institutional types. Second, multigroup analyses were conducted to determine if the within-group evidence of convergence and discrimination was consistent across institutional groupings.

Consistent with the recommendations of Byrne (1993), Marsh and Hocevar (1985), and Widaman (1985), confirmatory factor analysis was used to evaluate the multitrait-multimethod matrices for two- and four-year institutions and to assess the stability of relationships across the two types of institutions. The measured variables in the analyses consisted of the nine College BASE cluster scores and the nine self-report scales. In the first phase of this study, separate matrices of covariances among the measured variables for two- and four-year institutions were calculated and analyzed using the LISREL 8 computer program (Jöreskog and Sörbom, 1993). Because of significant multivariate skewness in the data, weighted least squares (i.e., asymptotically distribution-free) estimation procedures were employed for all of the analyses (Jöreskog and Sörbom, 1993). These methods were identical to those used by Pike (1994, 1995).

In order to evaluate convergence and discrimination within groups, five models were specified and tested. The first model contained six latent variables (i.e., factors). Two of the latent variables represented the different measurement methods, while the remaining four latent variables represented the subject area domains underlying College BASE and the self-report measures. The two latent variables representing the measurement methods were allowed to covary freely, as were the four latent variables representing outcome domains. Covariances

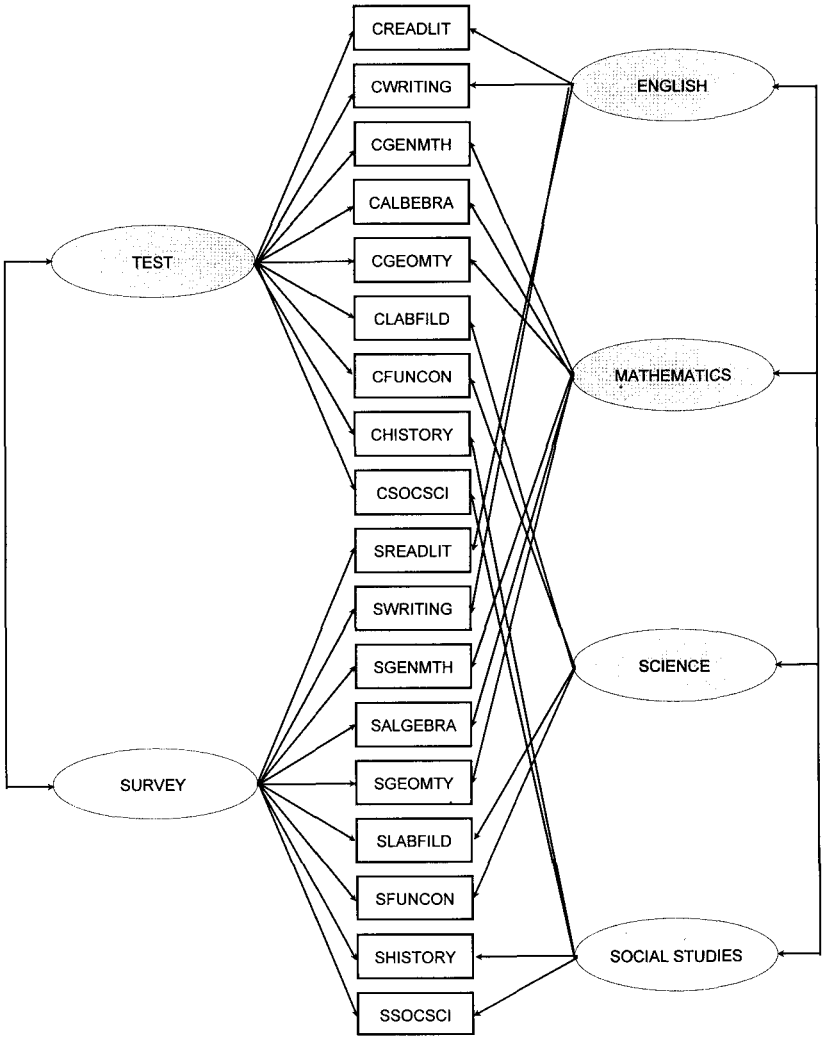

FIG. 1. Simplified baseline model for the multitrait-multimethod analyses. Uniquenesses have been omitted to improve readability.

between methods and trait factors were fixed at zero. A simplified version of the first model is provided in Figure 1.

The second model in the multitrait-multimethod analyses contained the two method factors, but not the four outcome factors. Consistent with the first model, the method factors were free to covary. A comparison of the goodness-of-fit statistics for the first and second models provided a test of the extent to

which the outcome domains were needed to explain relationships among the measured variables. This comparison represented an evaluation of the convergent validity of College BASE and self-report measures (Byrne, 1993).

The third and fourth models contained the six factors in the first (baseline) model. In the third model, however, the four latent variables representing the outcome domains were specified as being perfectly correlated. In the fourth model, the latent variables representing outcome domains were free to covary, but the two methods factor were perfectly correlated. Comparison of goodness-of-fit results for the first and third models provided a test of whether test scores and self-reports were able to discriminate among outcomes, with acceptance of the third model indicating that the two measurement methods did not discriminate among outcomes. Comparison of goodness-of-fit results for the first and fourth models provided a test of discrimination between measurement methods. Acceptance of the fourth model would imply that self-reports and test scores did not represent distinct measurement methods.

The fifth model was similar to the fourth confirmatory factor analysis model in that the latent variables representing outcome domains were free to covary, while the covariance between the latent variables representing measurement methods was constrained to a specific value. Unlike the fourth model, the covariance between measurement methods was fixed at zero. Although this model is not generally evaluated in multitrait-multimethod analysis, it was included to represent Pike's (1995) earlier findings. It is important to note that the selection of either the first or the fifth models would provide evidence of convergence and discrimination.

Byrne (1993) suggested that multitrait-multimethod factor models should be compared using traditional chi-square goodness-of-fit statistics and incremental fit indices. In this study, chi-square measures were used, but incremental fit indices were not used. The incremental fit indices were not used because asymptotically distribution-free estimation methods tend to produce inaccurate estimates of model fit for the null model (i.e., a model in which all observed variables are unrelated), and poor estimation of fit for the null model results in inaccurate and unstable incremental fit indices for the higher-order models tested in this study (Sugawara and MacCallum, 1993). As an alternative to reliance on incremental fit indices, Browne and Cudeck's (1989) cross-validation index, derived from the Akaike Information Criterion, was used in this study. This cross-validation index (CVI) has been shown to be appropriate when asymptotically distribution-free estimation methods are used and is robust with respect to departures from multivariate normality (Sugawara and MacCallum, 1993; Williams and Holahan, 1994).

Providing evidence of the stability of relationships across institutional types entailed establishing the invariance of the confirmatory factor analysis model across groups (Byrne, 1989; Jöreskog, 1971a; Marsh, 1994). In this phase of the

research, four models were specified and tested. The first model, with identical patterns of fixed and free parameters, but no constraints on the values of the free parameters, represented pattern invariance and provided the best possible multigroup model in terms of goodness of fit. Indeed, the chi-square value for the first model was equal to the sum of the chi-square values for the final models selected in the first phase of the research.

The second model used in the multigroup analyses was identical to the first model, with the added restriction that the values of the factor loadings were invariant across groups. The difference between the goodness-of-fit statistics for the first and second models represented a direct test of whether precisely the same constructs were being measured across two- and four-year institutions, with a nonsignificant chi-square difference providing evidence of measurement invariance across groups.

In the third model, factor loadings and covariances among the method and trait factors were constrained to be invariant across groups. A nonsignificant change in goodness of fit from the baseline to the third model provided evidence of measurement invariance and invariant relationships among methods and traits across institutions. Factor loadings, covariances, and uniquenesses were invariant across groups in the fourth model, indicating that these parameters were the same for both groups. The appropriateness of the four invariance models was assessed using traditional chi-square tests and the cross-validation index.

RESULTS

Within Groups

The results of the independent specification and testing of multitrait-multimethod models for two- and four-year institutions provided clear evidence of the convergent and discriminant validity of self-reports and test scores. Table 3 contains the goodness-of-fit results for these models.

Although the baseline model for two-year institutions produced a statistically significant value ($\chi^2 = 510.68$; $df = 110$; $p \leq .001$), the cross-validation index for this model was quite respectable (CVI = 0.77). The second model, in which there were no trait factors, produced a chi-square value of 1,013.19 ($df = 134$; $p \leq .001$). This value was significantly greater than the chi-square statistic for the baseline model ($\Delta\chi^2 = 502.51$; $\Delta df = 24$; $p \leq .001$), indicating that the trait factors were needed to explain the observed data. This interpretation also was supported by a relatively high cross-validation index (1.32). Most important, this finding provided clear support for the convergent validity of the test and self-report data for two-year institutions.

The third model was identical to the baseline model, except that it included the restriction that the four trait factors perfectly covary. The chi-square good-

TABLE 3. Goodness-of-Fit Results for the Within-Group Multitrait-Multimethod Analyses

	df	χ^2	Δdf	$\Delta\chi^2$	*CVI*
Two-Year Institutions					
[1] Baseline	110	510.68***	—	—	0.77
[2] No Traits	134	1013.19***	24	502.51***	1.32
[3] Perfectly Covarying Traits	116	700.48***	6	189.80***	0.98
[4] Perfectly Covarying Methods	111	588.48***	1	77.80***	0.86
[5] Unrelated Methods	111	510.96***	1	0.28	0.77
Four-Year Institutions					
[1] Baseline	110	470.15***	—	—	0.64
[2] No Traits	134	1111.55***	24	641.40***	1.50
[3] Perfectly Covarying Traits	116	652.82***	6	56.73***	0.88
[4] Perfectly Covarying Methods	111	526.88***	1	56.73***	0.71
[5] Unrelated Methods	111	469.87***	1	−0.28	0.64

*$p \leq 0.05$; **$p \leq 0.01$; ***$p \leq 0.001$

ness-of-fit result for this model was statistically significant and was significantly greater than the goodness-of-fit result for the baseline model ($\Delta\chi^2 = 189.80$; $\Delta df = 6$; $p \leq .001$). These results indicate that adding the restriction that traits perfectly covary significantly increased poorness of fit, thus providing evidence of discrimination among traits.

The fourth model was the baseline model with the restriction that methods perfectly covary. A comparison of goodness-of-fit results for the fourth and baseline models revealed that adding the restriction that methods perfectly covary significantly increased poorness of fit ($\Delta\chi^2 = 77.80$; $\Delta df = 1$; $p \leq .001$). This finding provided evidence of discrimination between methods.

The goodness-of-fit results for the fifth model were not significantly different from the results for the baseline model ($\Delta\chi^2 = 0.28$; $\Delta df = 1$; $p > .05$). Moreover, the fifth model and the baseline model had the same cross-validation index (0.77). Acceptance of the fifth model provided support for the convergent and discriminant validity of self-reports and test scores. However, it also implied that the two measurement methods were unrelated.

The results for four-year institutions, in the second half of Table 3, tell a similar story. The cross-validation index for the baseline model was a respectable 0.64, despite a significant chi-square value ($\chi^2 = 470.15$; $df = 110$; $p \leq .001$). Likewise, evidence of convergence was found in the fact that excluding trait factors from the model significantly increased poorness of fit ($\Delta\chi^2 = 641.40$; $\Delta df = 24$; $p \leq .001$). Requiring that the trait factors perfectly covary also significantly increased poorness of fit ($\Delta\chi^2 = 182.67$; $\Delta df = 6$; $p \leq .001$), as did requiring that methods perfectly covary ($\Delta\chi^2 = 56.73$; $df = 1$; $p \leq .001$).

Thus, the results for four-year institutions also provided evidence of the convergent and discriminant validity of self-reports and test scores.

Consistent with the results for two-year institutions, the four-year analyses revealed that requiring that the method factors be unrelated did not significantly increase poorness-of-fit ($\Delta\chi^2 = -0.28$; $df = 1$; $p > .05$).[2] The appropriateness of the fifth model was also supported by the fact that the cross-validation index was unaffected by imposing the restriction that the methods factors be unrelated.

Between Groups

Because the full multitrait-multimethod model with unrelated measurement methods (i.e., the fifth model) provided the most parsimonious acceptable explanation of the observed data for both two- and four-year institutions, it was used in the between-group analyses. Table 4 presents the goodness-of-fit results for the four models representing the various levels of between-group invariance.

Consistent with expectations, the chi-square value for Model 1, representing pattern invariance across groups, was the sum of the values from the within-group analyses ($\chi^2 = 980.83$; $df = 222$; $p \leq .001$). Despite the statistically significant chi-square value, the cross-validation index for the model was quite reasonable (0.70).

Adding the requirement that the factor loadings in the model be invariant across groups significantly increased poorness of fit ($\Delta\chi^2 = 165.41$; $df = 36$; $p \leq 0.1$). The cross-validation index for this model was only slightly higher than the CVI for the baseline model (0.79). Adding the restriction that the covariance among the traits be invariant across groups also significantly increased poorness of fit relative to the baseline model ($\Delta\chi^2 = 197.81$; $df = 42$; $p \leq .001$), as did adding the restriction that the uniquenesses be invariant across groups ($\Delta\chi^2 = 248.12$; $df = 60$; $p \leq .001$). For these last two models, the cross-validation indices were 0.80 and 0.82, respectively.

The results of the between-group analyses do not provide a definitive answer

TABLE 4. Goodness-of-Fit Results for the Between-Group Analyses

Model	*df*	χ^2	Δdf	$\Delta\chi^2$	*CVI*
[1] Baseline	222	980.93***	—	—	0.70
[2] Factor Loadings Invariant	258	1146.24***	36	165.41***	0.79
[3] Factor Loadings and Covariances Invariant	264	1178.64***	42	197.81***	0.80
[4] All Parameters Invariant	282	1228.95	60	248.12***	0.82

*$p \leq 0.05$; **$p \leq 0.01$; ***$p \leq 0.001$

to questions about the invariance of relationships across groups. On one hand, chi-square values suggest that, while the patterns of convergence and discrimination are invariant across groups, the factor loadings, factor covariances, and uniquenesses in the models are not invariant. On the other hand, cross-validation indices for all of the models in the between-group analyses were quite respectable. In an effort to better understand patterns of invariance across two- and four-year institutions, a detailed examination of the parameters in the model was undertaken.

Table 5 contains the two- and four-year common metric, completely standardized parameter estimates (i.e., factor loadings and uniquenesses) from the baseline model.[3] It also includes the corresponding factor loadings and uniquenesses for the fourth model in which all parameters were constrained to be invariant. It is important to realize that all parameter values in Table 5 are statistically significant. In the table, the parameter estimates for four-year institutions are in parentheses, while the parameter estimates for the fourth invariance model are italicized. To facilitate the identification of significant differences across groups, asterisks are included to identify those parameters, which when constrained to be invariant across groups, significantly added to poorness of model fit.

An examination of the common metric completely standardized factor loadings and uniquenesses, particularly for the model in which all parameters were invariant (i.e., the italicized factor loadings), reveals a pattern in which the test factor was more strongly related to College BASE scores than were the trait factors. The relative contributions of the traits and method factors was reversed for the self-report scales. The trait factors were more strongly related to self-reports than was the survey method factor. This pattern was most pronounced for four-year institutions, and it is within this context that group differences should be interpreted.

Relatively small differences between two- and four-year institutions were observed for factor loadings on the method factors. Modification indices revealed that constraining the test factor loadings for College BASE General Mathematics (CGENMATH) and for College BASE Lab and Field Techniques (CLABFIELD) contributed measurably to poorness of model fit. The only constrained survey method factor loading that contributed to poorness of fit was self-reports of geometry skills (SGEOMETRY). All three of these differences represent the clearest evidence that the trend for test scores to be most strongly related to the test factor and self-reports to be most strongly related to the survey factor is most prevalent among students attending four-year colleges and universities.

Differences in the trait factor loadings are equally subtle. No significant differences in trait factor loadings were found for the social studies trait and only one significant difference, College BASE Writing (CWRITING), was found for

TABLE 5. Common Metric Completely Standardized Factor Loadings and Uniquenesses

	Test	Survey	English	Math	Science	Social Studies	Uniqueness
College BASE							
CREADLIT	0.73		0.33				0.34
	(0.76)		(0.28)				(0.36)
	0.74		*0.32*				*0.34*
CWRITING	0.55		0.41				0.56
	(0.62)		(0.42)				(0.41)
	0.61		*0.41**				*0.46**
CGENMATH	0.67			0.43			0.28
	(0.73)			(0.50)			(0.30)
	0.71			*0.46*			*0.28*
CALGEBRA	0.50			0.53			0.43
	(0.43)			(0.69)			(0.38)
	0.49			*0.59*			*0.40*
CGEOMETRY	0.60			0.55			0.28
	(0.52)			(0.71)			(0.29)
	0.56			*0.62**			*0.27*
CLABFIELD	0.69				0.46		0.22
	(0.74)				(0.58)		(0.22)
	*0.72**				*0.51**		*0.22*
CFUNDCON	0.71				0.48		0.34
	(0.69)				(0.39)		(0.29)
	0.71				*0.42**		*0.31*
CHISTORY	0.65					0.57	0.25
	(0.68)					(0.48)	(0.31)
	0.68					*0.52*	*0.52*
CSOCSCI	0.71					0.50	0.22
	(0.68)					(0.59)	(0.16)
	0.71					*0.53*	*0.19*
Self-Reports							
SREADLIT		0.60	0.63				0.23
		(0.38)	(0.72)				(0.35)
		0.56	*0.61*				*0.29*
SWRITING		0.52	0.65				0.23
		(0.40)	(0.81)				(0.27)
		0.49	*0.69*				*0.24*
SGENMATH		0.52		0.61			0.26
		(0.34)		(0.80)			(0.36)
		0.43		*0.70*			*0.30*
SALGEBRA		0.35		0.71			0.25
		(0.15)		(0.96)			(0.20)
		0.24		*0.85**			*0.22*

TABLE 5.
(Continued)

Test	Survey	English	Math	Science	Social Studies	Uniqueness
SGEMOETRY	0.36		0.78			0.13
	(0.09)		(0.99)			(0.11)
	*0.24**		*0.89*			*0.13*
SLABFIELD	0.49			0.70		0.25
	(0.25)			(0.84)		(0.26)
	0.39			*0.75*		*0.26*
SFUNDCON	0.57			0.60		0.25
	(0.37)			(0.79)		(0.30)
	0.49			*0.66**		*0.28**
SHISTORY	0.70				0.41	0.40
	(0.62)				(0.39)	(0.40)
	0.67				*0.36*	*0.37*
SSOCSCI	0.92				0.16	0.11
	(0.91)				(0.21)	(0.16)
	0.85				*0.17*	*0.19*

the English trait. However, several significant differences were observed for the mathematics and science traits. A comparison of the magnitudes of trait factor loadings across two- and four-year institutions reveals that the trait factor loadings for four-year institutions generally were larger than the trait factor loadings for two-year institutions. It also may be significant that all but two of the between-group differences (CWRITING and SFUNDCON) occurred for observed measures in which there was significant heterogeneity of variance across type of institution.

It is also important to note that the uniqueness parameters in the multitrait-multimethod models were generally stable across the two types of institutions. An examination of the modification indices for the uniquenesses revealed that only the uniqueness parameter for the College BASE writing scale (CWRITING) significantly added to poorness of fit when it was constrained to be invariant across groups. The stability in uniquenesses across groups provides clear evidence that the overall explanatory power of the multitrait-multimethod model was essentially the same across groups.

Table 6 contains the common metric completely standardized covariances among the method and trait factors. As with the results in Table 5, parameter estimates for four-year institutions are in parentheses, while the parameter estimates for the model representing total invariance are in italics. An examination of the parameter estimates in Table 6 reveals that the covariances among the trait factors were generally larger for two-year than for four-year institutions. In

TABLE 6. Common Metric Completely Standardized Covariances Among the Facotrs

	Test	Survey	English	Math	Science	Social Studies
Test	1.00					
	(1.00)					
	1.00					
Survey		1.00				
		(1.00)				
		1.00				
English			1.00			
			(1.00)			
			1.00			
Mathematics			0.42	1.00		
			(0.29)	(1.00)		
			0.38	*1.00*		
Science			0.51	0.84	1.00	
			(0.26)	(0.89)	(1.00)	
			0.40	*0.86*	*1.00*	
Social Studies			0.54	0.52	0.64	1.00
			(0.19)	(0.59)	(0.52)	(1.00)
			*0.40**	*0.58**	*0.59**	*1.00*

addition, modification indices indicate that constraining any of the correlations between the social studies factor and the other three trait factors to be invariant measurably added to poorness of fit. This finding is most interesting given the fact that the factor loadings on the social studies factor were stable across two- and four-year institutions.

DISCUSSION

Obviously the generalizability of the findings from the present research limited in terms of the institutions and the measure used in the study. Additional research is needed with larger, more diverse samples of institutions, and additional research is needed with a variety of educational outcome measures. Despite these limitations, the results of the present research provide some important information about the validity of using self-reports of cognitive development during college as proxies for test scores in a national assessment of college student outcomes. The findings of this study can be summarized as follows:

1. The within-group multitrait-multimethod analyses provided clear support for the convergence of self-reports and test scores. Goodness-of-fit tests indi-

cated that four outcome domains and two method factors underlie the relationships among self-reports and test scores.
2. Likewise, the within-group analyses found evidence of discrimination among the four trait factors and between the two method factors. In fact, the within-group analyses suggested that the two method factors were unrelated. This finding was in sharp contrast to previous research indicating a moderate positive correlation between method factors.
3. The results of the between-group analyses were ambiguous. On one hand, the chi-square goodness-of-fit tests suggested that, while the general pattern of convergence and discrimination was the same across two- and four-year institutions, the strength of the relationships between observed measures and the method and trait factors differed by type of institution. On the other hand, cross-validation indices and parameter estimates indicated that differences by type of institution were relatively subtle, representing differences in the magnitudes of relationships, not differences in the nature of the relationships.

These findings have several implications for the use of self-reports as proxies for test scores, the most obvious implication being the nature of the relationship between self-reports and test scores both between and within groups. Jöreskog (1971b) described three levels of equivalence among different measures. He termed the most basic level of equivalence, *congeneric tests*. This level of equivalence occurs when several measures all represent the same construct. At the next level in Jöreskog's hierarchy are *tau-equivalent tests*, in which the factor loadings of different measures of the same contruct are all identical. That is, each observed variable contributes equally to the construct. The highest level in Jöreskog's hierarchy is represented by *parallel tests*, in which both the factor loadings and uniquenesses for the observed measures are identical.

The factor loadings in Table 5 strongly suggest that, within groups, the observed measures of the four outcome domains are congeneric tests. While these observed measures are significantly related to one, and only one, outcome domain, the strength of the relationships differs across the measures. For example, self-reports and test scores for Reading and Literature and Writing are all significantly related to the outcome domain titled "English." However, the strength of the relationships differs significantly.

Jöreskog's hierarchy is also useful in defining the nature of the relationships between self-reports and test scores between groups. The presence of pattern invariance across groups is evidence that the observed measures are congeneric across groups. That is, observed measures represent the same general constructs, but the strength of those relationships may not be precisely the same for different groups. The next level in the hierarchy, tau-equivalence across groups, occurs when the contributions of observed measures to the trait factors are

identical for different groups. At the apex of the hierarchy is the presence of parallel tests across groups. For observed measures from different groups to be considered parallel tests, all measurement parameters should be invariant across groups.

The data clearly show that self-reports and test scores are, at least, congeneric measures between, as well as within, groups. Goodness-of-fit tests clearly support the appropriateness of pattern invariance across groups, and the factor loadings in Table 5 provide additional evidence that patterns of factor loadings are the same for both two- and four-year institutions. What is unclear is whether the actual parameter values for the two groups can be said to come from the same or different populations. They may even be parallel tests between groups.

An inspection of the parameter estimates contained in Table 5 suggests that differences in factor loadings are subtle, reflecting the fact that, for four-year institutions as compared to two-year institutions, there is a greater tendency for self-report items to be more strongly related to the trait factors than to the survey method factor. For four-year institutions, the test method factor was more strongly related to College BASE scores than were the four trait factors. If these differences are significant, then what is meant by English, mathematics, science, and social studies outcomes is not quite the same for two- and four-year institutions, and comparisons of results across institutions could be misleading.

One surprising finding of the present research was that the two method factors were unrelated for both two- and four-year institutions. Previous research (Pike, 1994) had shown moderate positive relationships between the measurement factors when the groups were combined. This seeming inconsistency can be explained as a statistical artifact. As reported in Table 2, four-year college means on the College BASE scales, and to a lesser extent the self-report scales, were significantly higher than those for two-year colleges. When the groups are combined, consistent mean differences between two- and four-year colleges and universities introduce spurious covariance into the relationship between orthogonal (unrelated) measures.

While the inconsistency in findings is explainable in statistical terms, the absence of a moderate positive relationship between measurement methods is troubling from a policy standpoint. As Pike (1994) noted, unrelated method factors create method-specific variance in the observed variables and attenuate the relationships among observed measures. A direct consequence of the attenuation of relationships among observed variables is that the observed variables will be relatively poor representations of the same educational outcomes domain, and simple comparisons of actual test scores and self-reports will be misleading. Obviously it is possible to disattenuate these relationships using the statistical methods incorporated in the present research. However, these statisti-

cal methods are not easily explainable to a lay public and may lack the necessary credibility for use in a national assessment of college student learning.

The results of the present research also have important implications from a statistical and methodological standpoint. The multitrait-multimethod models used in this study are extremely complex and difficult to estimate, particularly since the observed data do not have a multivariate normal distribution. One practical consequence of the complexity in the present research is that parameter estimates may converge toward local minima (i.e., parameter values that satisfy the converge criteria, but do not represent the optimal explanation of the observed data). Evidence for a local minimum can be found in the fact that, for four-year colleges and universities, the baseline model did not provide as good an explanation of the data as did a more restricted model.

A second, more vexing, problem with model complexity and data distribution was the inability to identify an appropriate null (worst fitting) model for the between-group analyses. In the present research, the null model actually represented the condition of parallel tests while the baseline model represented the condition of congeneric tests. Differences between the models provided an indication of the poorness of fit created by moving from the assumption that measures were congeneric across groups to the assumption that measures were parallel across groups. What could not be ascertained was whether this additional poorness of fit was significant in the large scheme. Until research utilizing more normally distributed data and more restrictive null models is available, it will be impossible to adequately test whether observed measures are parallel across groups.

CONCLUSIONS

Can self-reports of student learning and academic development serve as proxies for more traditional measures of student achievement? The answer is still a cautious "yes." For both two- and four-year colleges and universities, self-reports and test scores based on the same set of specifications do represent the same educational outcome domains (i.e., they are congeneric). However, there is not a one-to-one correspondence between self-reports and more objective measures of achievement. Consequently, using self-reports as general indicators of achievement can be justified, but substituting specific self-reports for test scores cannot be justified based on the results of the present research.

For policymakers and researchers interested in examining results for educational outcome domains across groups, the same caveat holds true. Educational outcome domains may represent congeneric, not parallel, measures across groups. Individuals are justified in assuming that what is generally contained in the domain of English outcomes is similar across two- and four-year institutions. However, the English domains of the two groups may not be precisely

the same and comparisons across different types of institutions may lead to erroneous conclusions.

Just as many of the participants in the first NCES study design workshop concluded that developing a national test of college student achievement would be a difficult and expensive task, so too will the development of a national survey of college student achievement be a difficult, if not expensive, task. First and foremost, there must be a consensus regarding what are critical thinking and communicating and what are key indicators of those traits. Despite the efforts of the members of the second study design workshop, Banta (1991) is correct in arguing that we have yet to reach agreement on what are the components of critical thinking and communicating.

Once agreement about what is to be measured has been achieved, the problems identified in this research concerning the comparability of measurements will have to be addressed. This will not be an easy task. Moreover, using sophisticated statistical procedures to represent common outcome domains may not be credible to an American public that does not presently trust higher education. A very real danger is that a public hungry for simple answers to complex questions will forget that self-reports of learning and academic development are not precisely the same as more traditional measures of the same outcomes, and draw erroneous conclusions about the quality and effectiveness of postsecondary education.

NOTES

1. The term *test* is used broadly to refer to a variety of objective and subjective measures of student achievement, including multiple-choice examinations and performance assessments.
2. The negative chi-square change statistics is probably the result of the estimation procedure converging to a local minimum that satisfied the convergence criterion, but did not provide the optimal representation of the observed data.
3. Common metric completely standardized parameter estimates are obtained by standardizing both the observed and the latent variables. The observed and latent variables are rescaled so that the weighted average of the group covariance matrices is a correlation matrix. The common metric approach produces identical standardized estimates for parameters constrained to be equal across groups. It also allows for direct comparisons across groups of freely varying parameters (Jöreskog and Sörbom, 1993).

REFERENCES

Anaya, G. (1992). Cognitive development among college undergraduates. Unpublished doctoral dissertation, University of California, Los Angeles.

Armor, D. J. (1974). Theta reliability and factor scaling. In H. L. Costner (ed.), *Sociological Methodology 1973–1974* (pp. 17–50). San Francisco: Jossey-Bass.

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

Baird, L. L. (1976). *Using Self Reports to Predict Student Performance.* New York: College Entrance Examination Board.

Banta, T. W. (1991). Toward a plan for using national assessment to ensure continuous improvement in higher education. Unpublished manuscript, Center for Assessment Research and Development, Knoxville, TN. ERIC Document Reproduction Service No. ED 340 753.

Berdie, R. F. (1971). Self-claimed and tested knowledge. *Educational and Psychological Measurement* 31: 629–636.

Browne, M. W., and Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research* 24: 445–455.

Byrne, B. M. (1989). Multigroup comparisons and the assumption of equivalent construct validity across groups: Methodological and substantive issues. *Multivariate Behavioral Research* 24: 503–523.

Byrne, B. M. (1993). *Structural Equation Modeling with EQS and EQS/Windows*. Thousand Oaks, CA: Sage.

Campbell, D. T., and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* 56: 81–105.

College Entrance Examination Board (1983). *Academic Preparation for College: What Examinees Need to Know and Be Able to Do*. New York: Author.

Cronbach, L. J., and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin* 52: 281–302.

Daly, J. A. (1994). Assessing speaking and listening: Preliminary considerations for a national assessment. In A. Greenwood (ed.), *The National Assessment of College Student Learning: Identification of the Skills to Be Taught, Learned, and Assessed* (pp. 113–161). Washington, DC: U.S. Government Printing Office, NCES 94-286.

Dumont, R. G., and Troelstrup, R. L. (1980). Exploring relationships between objective and subjective measures of instructional outcomes. *Research in Higher Education* 12: 37–51.

Dunbar, S. (1991). *On the Development of a National Assessment of College Student Learning: Measurement Policy and Practice in Perspective*. University of Iowa, Iowa City, IA. ERIC Document Reproduction Service No. ED 340 755.

Elliott, E. (1991). Charge to participants. In A. Greenwood (ed.), *National Assessment of College Student Learning: Issues and Concerns* (pp. 24–31). Washington, DC: U.S. Government Printing Office.

Ewell, P. T. (1991). To capture the ineffable: New forms of assessment in higher education. In G. Grant (ed.), *Review of Research in Education* (vol. 17). Washington, DC: American Educational Research Association.

Ewell, P. T. (1994). A matter of integrity: Accountability and the future of self-regulation. *Change* 26: 25–29.

Ewell, P. T., Lovell, C. D., Dressler, P., and Jones, D. P. (1994). *A Preliminary Study of the Feasibility and Utility for National Policy of Instructional "Good Practice" Indicators in Undergraduate Education*. Washington, DC: U.S. Government Printing Office. NCES 94-437.

Fiske, D. W. (1982). Convergent-discriminant validation of measurements in research strategies. In D. Brinberg and L. Kidder (eds.), *Forms of Validity in Research* (New Directions for the Methodology of Social and Behavioral Science Series, No. 12, pp. 77–92). San Francisco: Jossey-Bass.

Halpern, D. F. (1994). A national assessment of critical thinking skills in adults: Taking steps toward the goal. In A. Greenwood (ed.), *The National Assessment of College Student Learning: Identification of the Skills to Be Taught, Learned, and Assessed* (pp. 24–64). Washington, DC: U.S. Government Printing Office. NCES 94-286.

House, E. R. (1993). *Professional Evaluation: Social Impact and Political Consequences*. Newbury Park, CA: Sage.

Jöreskog, K. G. (1971a). Simultaneous factor analysis in several populations. *Psychometrika* 35: 409–426.

Jöreskog, K. G. (1971b). Statistical analysis of sets of congeneric tests. *Psychometrika* 36: 109–133.

Jöreskog, K. G., and Sörbom, D. (1993). *LISREL 8*. Chicago: Scientific Software.

Marsh, H. W. (1994). Confirmatory factor analysis models of factorial invariance: A multifaceted approach. *Structural Equation Modeling* 1: 5–34.

Marsh, H. W., and Hocevar, D. (1985). The application of confirmatory factor analysis to the study of self concept: First and higher order factor structures and their invariance across age groups. *Psychological Bulletin* 97: 562–582.

McClenney, K. (1993). Assessment in an era of empowerment. *Assessment Update: Progress, Trends, and Practices in Higher Education* 5(1): 1–2, 4–6.

National Education Goals Panel Resource Group on Adult Literacy and Lifelong Learning (1991). Adult literacy and lifelong learning. In National Education Goals Panel, *Measuring Progress Toward the National Education Goals: Potential Indicators and Measurement Strategies* (pp. 81–98). Washington, DC: U.S. Government Printing Office.

Osterlind, S. J. (1989). *College BASE: Guide to Test Content*. Chicago: Riverside.

Osterlind, S. J., and Merz, W. R. (1992). *College BASE Technical Manual*. University of Missouri–Columbia: Center for Educational Assessment.

Pace, C. R. (1987). *CSEQ Test Manual and Norms*. Los Angeles: Center for the Study of Evaluation.

Pascarella, E. T., and Terenzini, P. T. (1991). *How College Affects Students: Findings and Insights from Twenty Years of Research*. San Francisco: Jossey-Bass.

Perkins, D., Jay, E., and Tishman, S. (1994). Assessing thinking: A framework for measuring critical thinking and problem-solving skills at the college level. In A. Greenwood (ed.), *The National Assessment of College Student Learning: Identification of the Skills to Be Taught, Learned, and Assessed* (pp. 65–111). Washington, DC: U.S. Government Printing Office. NCES 94-286.

Pike, G. R. (1992a). *A Generalizability Analysis of the College Basic Academic Subjects Examination*. Knoxville, TN: Center for Assessment Research and Development, University of Tennessee.

Pike, G. R. (1992b). The components of construct validity: A comparison of two measures of general education outcomes. *Journal of General Education* 41: 130–159.

Pike, G. R. (1994, November). The relationship between self-report and objective measures of student achievement. Paper presented at the annual meeting of the Association for the Study of Higher Education, Tucson, AZ.

Pike, G. R. (1995). The relationship between self reports of college experiences and achievement test scores. *Research in Higher Education* 36: 1–22.

Pohlmann, J. T., and Beggs, D. L. (1974). A study of the validity of self-reported measures of academic growth. *Journal of Educational Measurement* 11: 115–119.

Ratcliff, J. L. (1991). *What Type of National Assessment Fits American Higher Education?* National Center for Teaching, Learning, and Assessment, Pennsylvania State University, University Park, PA. ERIC Document Reproduction Service No. ED 340 763.

Schmitt, N., and Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement* 10: 1–22.

Sugawara, N., and MacCallum, R. C. (1993). Effect of estimation method on incremental fit indexes for covariance structure models. *Applied Psychological Measurement* 17: 365–378.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement* 9: 1–26.

Williams, L. J., and Holahan, P. J. (1994). Parsimony-based fit indices for multiple-indicator models: Do they work? *Structural Equation Modeling* 1: 161–189.

Wingspread Group on Higher Education (1993). *An American Imperative: Higher Expectations for Higher Education.* The Johnson Foundation.