DOCUMENT RESUME

ED 373 632 HE 027 618

AUTHOR Pike, Gary R.
TITLE The Relationship between Self Reports of College Experiences and Achievement-Test Scores. AIR 1994 Annual Forum Paper.
PUB DATE May 94
NOTE 38p.; Paper presented at the Annual Forum of the Association for Institutional Research (34th, New Orleans, LA, May 29-June 1 1994).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; *College Outcomes Assessment; College Seniors; *Comparative Analysis; Higher Education; *Self Evaluation (Individuals); Student Attitudes; Test Format; Testing; Test Reliability
IDENTIFIERS *AIR Forum; *College Basic Academic Subjects Examination; *Student Self Report; University of Tennessee Knoxville

ABSTRACT

This paper examines the proposed use of student self-report data as proxies for College Basic Academic Subjects Examination (College BASE) scores and as policy indicators of good educational practice. A recent study by the National Center for Higher Education Management Systems had recommended this use of student self-reports. For this study 540 seniors at the University of Tennessee, Knoxville (UTK) completed the College BASE and a survey designed to elicit students' perceptions of their college experiences. Also, a randomly selected group took the College BASE, and those who did not take the test wrote a series of essays and completed a senior survey. Analysis using a multitrait-multimethod approach provided limited support for using self reports of cognitive development during college as proxies for College BASE test results. Careful examination of factor loadings revealed strong evidence of convergence only for the mathematics domain. Convergence in English and science was relatively weak. However the use of self reports of college experiences as policy indicators to guide institutions in improving the quality of undergraduate education was supported, particularly in mathematics and science. Includes three tables and one figure. Contains 46 references. (JB)

THE RELATIONSHIP BETWEEN SELF REPORTS OF COLLEGE EXPERIENCES AND ACHIEVEMENT-TEST SCORES

Gary R. Pike
Senior Research Analyst

CENTER FOR EDUCATIONAL ASSESSMENT
UNIVERSITY OF MISSOURI–COLUMBIA
100 TOWNSEND HALL
COLUMBIA, MO 65211
314/882-2963

for Management Research, Policy Analysis, and Planning

This paper was presented at the Thirty-Fourth Annual Forum of the Association for Institutional Research held at The New Orleans Marriott, New Orleans, Louisiana, May 29, 1994 - June 1, 1994. This paper was reviewed by the AIR Forum Publications Committee and was judged to be of high quality and of interest to others concerned with the research of higher education. It has therefore been selected to be included in the ERIC Collection of Forum Papers.

Jean Endo
Editor
Forum Publications

Abstract

The recently enacted national education goals represent a renewed effort by the federal government to promote outcomes assessment. In fact, as part of national goal 5.5, the federal government has proposed implementing a national test of undergraduate education outcomes. Several participants in a federally sponsored workshop on implementation recommended that a national test be a long-term goal, and that self reports of students' college experiences should be one of several indicators used in the interim. This study evaluated the use of students' self reports both as proxies and policy indicators for a national assessment. Results indicated that self reports of cognitive development during college should be used with care as proxies for a national test. Using students' reports of their experiences during college as policy indicators to guide institutions in improving the quality of undergraduate education was supported by the results of the present study..

THE RELATIONSHIP BETWEEN SELF REPORTS OF COLLEGE EXPERIENCES AND ACHIEVEMENT-TEST SCORES

The last half of the 1980s can be characterized as a period of growing public dissatisfaction with American higher education. During this time, articles began appearing in the popular press criticizing colleges and universities for increasing costs and decreasing levels of student achievement (Grossman, 1988; Hartle, 1986). Several blue-ribbon advisory panels also issued reports criticizing the quality and effectiveness of higher education (Association of American Colleges, 1985; Boyer, 1987; National Governors' Association, 1986; National Institute of Education Study Group on the Conditions of Excellence in American Higher Education, 1984). While these reports differed in tone and emphasis, all of the advisory panels recommended that colleges and universities set challenging goals for student achievement and then assess student progress toward those goals (Ewell, 1991).

As a result of public calls for assessment and accountability, states, the federal government, and accrediting associations have taken actions designed to encourage colleges and universities to adopt comprehensive assessment and evaluation programs (House, 1993; Pike, 1992). By 1990, surveys by the American Council on Education reported that more than 80 percent of all public and private colleges and universities had implemented, or were in the process of implementing, assessment programs (El-Khawas, 1990).

Despite the increased use of outcomes assessment for accountability and improvement, criticisms of higher education continue. Typical of these criticisms is the recent report of the

Wingspread Group on Higher Education. This report called on colleges and universities to redouble their assessment efforts.

The recently enacted national education goals represent another effort to promote the assessment of student outcomes. National education goal 5.5 states: By the year 2000, "the proportion of college graduates who demonstrate an advanced ability to think critically, communicate effectively, and solve problems will increase substantially" (National Education Goals Panel Resource Group on Adult Literacy and Lifelong Learning, 1991, p. 81). In order to monitor progress toward that goal, the Department of Education has proposed that a test similar to the *National Assessment of Educational Progress* (*NAEP*) be developed and administered to college students (Elliott, 1992).

Several participants in a federally sponsored workshop on implementing a national test of college-student achievement voiced reservations about the feasibility of developing and implementing a national assessment (Ewell, Lovell, Dressler, and Jones, 1993). Banta (1991), for example, raised questions about the desirability and practicality of achieving a national consensus on the outcomes to be assessed, while Dunbar (1991) identified several potential technical problems with creating a test that would be reliable and valid as a national assessment. Other participants were more optimistic. Ratcliff (1991) concluded that a national assessment program is feasible. However, he urged that the development of a national assessment be a long-term project. He suggested that alternative measures be used in the interim as proxies for a national test and as a guide for policy decisions.

In 1991, the National Education Goals Panel Resource Group on Adult Literacy and Lifelong Learning also recommended that alternatives to a national test be considered. The group

argued that alternative measures of cognitive development during college could serve as proxies for a national achievement test. The Resource Group also suggested that measures of good practice in postsecondary education be used to supplement the results of a national test, providing data about college experiences that could be used in educational policy making.

Based on the results of the workshop and the Resource Group report, the National Center for Educational Statistics (NCES) contracted with the National Center for Higher Education Management Systems (NCHEMS) to conduct a preliminary study of the feasibility of using measures of good practice as indicators of the quality and effectiveness of undergraduate education. NCHEMS evaluated previous research on a variety of possible indicators, including institutions' general education requirements, reliance on active-learning methods in classroom teaching, and students' reports of their college experiences. One of the conclusions of the NCHEMS report was that self-report data on students' college experiences have "moderate" to "high" potential as proxies for a national test and as policy indicators of good educational practice.

This paper reexamines the conclusions of the NCHEMS report in light of recent data on the relationship between self reports of college experiences and scores on a widely used test of general-education knowledge and skills, the *College Basic Academic Subjects Examination* (*College BASE*). Initially, this research examines the feasibility of using self reports of cognitive development during college as proxies for *College BASE* scores. The study then examines the relationship between *College BASE* scores and self reports of college experiences (e.g., involvement and the college environment) in order to determine whether self reports can be used as policy indicators to improve educational practice and enhance student performance.

Previous Research

Self Reports as Proxies

A variety of studies have found moderate correlations between self reports of cognitive development and achievement test scores. Generally, the strongest correlations have been found in studies designed to validate self reports (Baird, 1976b; Berdie, 1971; Pohlmann and Beggs, 1974). Berdie (1971), for example, found correlations ranging from 0.47 to 0.74 between self reports and test scores on a list of famous people, while Pohlmann and Beggs (1974) reported correlations of 0.52 to 0.67 for course outcomes.

Studies of the relationships between self reports and tests of college outcomes have also reported moderate correlations, although these correlations are generally lower than the correlations reported in validity studies (Anaya, 1992; Astin, 1993; Dumont and Troelstrup, 1980). Dumont and Troelstrup (1980) found correlations of 0.21 to 0.24 between self reports and scores on the *College Outcome Measures Program* (*COMP*) examination, while Astin (1993) reported low to moderate correlations between self reports of growth in knowledge and scores on tests such as the *Graduate Record Examination* (*GRE*) and *National Teachers Examination* (*NTE*).

The results of these studies suggest that two factors can influence the strength of the relationship between self reports of cognitive development and achievement test scores. First, content correspondence seems to be positively related to the correlation between self reports and test scores. In studies, such as those by Berdie (1971) and Pohlmann and Beggs (1974), in which the contents of the self reports and achievement tests were very similar, correlations were

moderate to high. In studies where there was less content overlap between self reports and tests the correlations were much lower.

Another factor that may serve to reduce the size of the correlations between self reports of cognitive development and achievement test scores is differences in the measurement methods. Astin (1993) noted that standardized achievement tests tend to have high fidelity, but narrow band width. That is, standardized tests tend to measure achievement very accurately over a relatively narrow range. In contrast, self reports tend to have lower fidelity, but greater band width. In other words, self reports measure a broad array of outcomes, but they are not as precise as standardized tests. These differences give rise to method-specific variance in scores, thereby attenuating the correlations between test scores and self reports.

Self Reports as Policy Indicators

Many of the studies designed to evaluate the feasibility of using students' self reports of their college experiences as indicators of good educational practice have focused on the relationship between self reports of college experiences and self reports of cognitive gains. These studies have found that self reports of cognitive gains are strongly related to students' perceptions of the college environment (Baird, 1976a; Friedlander, 1980; Pace, 1990), quality of student effort (Friedlander, 1991; Pace, 1987; Porter, 1982), and students' relationships with faculty and peers (Pascarella and Terenzini, 1978; Terenzini, Pascarella, and Lorang, 1982; Terenzini and Wright, 1987).

Studies examining the relationship between self reports of college experiences and standardized test scores have been relatively rare, and the relationships between college

experiences and test scores has been much smaller in magnitude than relationships between college experiences and self reports of cognitive development (Astin, 1993; Friedlander, 1980). Moreover, some studies, such as the joint factor analysis conducted by Davis and Murrell (1990), have failed to detect a significant relationship between self reports of college experiences and achievement-test scores. When significant relationship have been detected, the greatest effects have been for environmental variables, student involvement, student-faculty interaction, and peer relations (Astin, 1993).

While previous research suggests that self reports can serve as policy indicators and, perhaps, as proxies for standardized achievement tests, no studies have directly addressed this issue within the context of the current outcomes assessment movement. Given the interest in developing a national test of educational progress and indicators of good educational practice, it is important to know under what circumstances and with what confidence self reports of cognitive development can be used as proxies for achievement-test results. It is also important to know the extent to which student reports of college experiences are related to test scores in order to identify those aspects of educational practice that can be modified in order to enhance student achievement.

Criteria for Evaluating Self Reports

In their review, NCHEMS staff enumerated four criteria for evaluating reports of cognitive development and educational practice: (1) self reports should represent broad-based outcomes or educational practices; (2) the measures should reliably covary with test scores; (3) the relationships between self reports and test scores should persist across different educational

settings; and (4) the self-report measures should represent significant phenomena that can be used to inform policy actions (Ewell, Lovell, Dressler, and Jones, 1993). Importantly, all of the measures of cognitive development and educational practice included in the present study met the four NCHEMS criteria. However, these criteria alone are not sufficient. Additional information about construct and criterion-related validity is required before the use of self reports as proxies and policy indicators can be considered valid.

Self Reports as Proxies

In order to serve as proxies for achievement-test scores, self reports must measure the same construct as the achievement test. If both instruments measure the same constructs, self reports can be substituted for test scores using appropriate statistical adjustments for differences in errors of measurement (Jöreskog, 1971). However, if both instruments do not measure the same construct, no amount of statistical manipulation will allow self reports of cognitive development to serve as proxies for test scores. Consequently, the *convergent validity* of self reports and test scores is a key criterion for determining whether self reports can be used as proxies for achievement-test data (Cronbach and Meehl, 1955).

A variety of techniques can be used to evaluate the convergent validity of two sets of measures such as self reports and test scores (Widaman, 1985). Of the available approaches, the analysis of multitrait-multimethod matrices represents an extremely popular and powerful tool (Campbell and Fiske, 1959). The primary strength of the multitrait-multimethod approach is that it allows a researcher to assess the strength of the *true* relationship between two or more instruments or methods (convergence), while simultaneously providing a measure of whether the

different instruments/methods can differentiate among constructs (discrimination) (Schmitt and Stults, 1986; Widaman, 1985).

According to Campbell and Fiske (1959), multitrait-multimethod analysis requires that two or more traits (constructs) be measured using two or more methods. Significant correlations among different methods of measuring the same trait or construct provide evidence of convergence, while the absence of significant correlations among traits provides evidence of discrimination. Empirical research on multitrait-multimethod matrices reveals 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 indicating convergence and discrimination.

In the present research convergence and discrimination were assessed using four factor-analytic models. By comparing these four models it is possible to test statistically the relative strength of the relationships indicating converge and discrimination. In addition, using these models it is possible to partition the variability in students scores into trait-specific and method-specific variance. This approach is described in the methods section of this paper.

Self Reports as Policy Indicators

Two criteria form the basis for evaluating the appropriateness of using self reports of college experiences as policy indicators. The first criteria is that the self reports should be significantly related to achievement-test scores. The second criteria is that these relationships should be substantively meaningful. Obviously, if self reports of college experiences are not related to test

scores the self reports can have little practical utiiity for improving student achievement. However, statistical significance is a function of sample size, and it is possible to find statistically significant relationships that do not represent substantively meaningful relationships. Thus, a second important question to be asked concerning the utility of self reports is whether changes in students' college experiences have a meaningful impact on test scores. Again, the specific techniques used to evaluate the strength and significance of relationships between college experiences and test scores will be examined in detail in the discussion of the research methods used in this study.

Research Methods

Subjects

The subjects for this study were 540 seniors at the University of Tennessee, Knoxville (UTK) who completed *College BASE* and a survey designed to elicit students' perceptions of their college experiences. The seniors who were tested using *College BASE* were randomly selected from the population of seniors at UTK, and those seniors who did not take *College BASE* were required to write a series of essays and complete the senior survey.

Approximately 52 percent of the seniors included in this research were females and 48 percent were males. Slightly less than 95 percent of the subjects were white, 4 percent were Black, and Asian, hispanic, and other racial/ethnic categories comprised slightly more than 1 percent of the sample. The average entering *ACT Assessment* composite score for the sample was 24.7, the mean cumulative grade point average was 3.10, and the average *College BASE* score was 320. Analysis of variance and chi-square tests revealed that the students included in

the present research were not substantively different from other seniors at UTK in terms of their background characteristics (i.e., gender, race, and ACT scores), cumulative grade point average, or responses to items on the college experiences survey.

Instruments

As indicated above, the data for this study were obtained from the *College Basic Academic Subjects Examination* and the senior survey. In this study, *College BASE* scores served as the criterion variables against which self reports were judged. *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: English, mathematics, science, and social studies. Subject scores are built upon content clusters which, in turn, are based on specific skills. For example, English scores are based on two content clusters: reading and literature and writing. The cluster score for reading and literature is based on skills related to reading analytically, reading critically, and understanding literature. In addition to a composite (total) score, numerical scores are provided for each subject and cluster. Ratings of high, medium, or low are provided for each skill (Osterlind, 1989). In this study, the nine cluster scores for the exam were used to represent the four subject areas. Reliability estimates for the cluster scores range from a low of 0.67 for writing to a high of 0.84 for algebra (Pike, 1992).

Measures of cognitive development during college and students' perceptions of their college experiences were all obtained from the senior survey. In this study, four cognitive-development scales were used. These scales represented the domains of English, mathematics, science, and

social studies. Items included in these scales were selected because of their correspondence to the content clusters measured by *College BASE*. Content correspondence was greatest for the mathematics domain and lowest for the social studies domain. Reliability estimates for the four cognitive-development scales ranged from 0.63 for English to 0.74 for social studies.

The remainder of the survey yielded 14 scales representing students' perceptions of the college environment, involvement, and interactions with faculty and peers. Two scales, the presence of a supportive environment for academic achievement and the presence of a supportive environment for creativity and artistic achievement represented the environment measures. Reliability estimates for both scales were 0.73.

Ten scales represented four distinct aspects of student involvement. Academic involvement was measured by three scales: use of the library, overall class effort, and academic effort outside the classroom. Reliability estimates were 0.83, 0.82, and 0.79, respectively. Involvement in writing was measured by two scales representing the amount of writing done for classes and the extent to which students sought feedback about their writing. Reliability estimates were 0.77 and 0.79. Extracurricular involvement also was represented by two scales: involvement in intramurals and recreation activities and involvement in clubs and organizations. The reliability estimates for the two scales were 0.86 and 0.89. Scales for involvement in cultural activities, specifically involvement in art music and theatre, produced the lowest reliability estimates, 0.70, 0.60, and 0.57, respectively.

Student interaction with faculty and peers was represented by two scales. These scales were modified versions of the scales developed by Pascarella and Terenzini (1978). The reliability

estimate for the faculty-student interaction scale was 0.92, while the reliability estimate for the peer-interaction scale was 0.84.

Data Analysis

The data analysis was a two-step process. First, a multitrait-multimethod analysis was performed to evaluate whether student reports of cognitive development during college could be used as proxies for *College BASE* scores. Second, a structural equation model was used to determine whether student reports could be used as policy indicators to improve performance on the *College BASE*. Both analyses made use of covariance structure modeling (CSM).

Consistent with the recommendations of Byrne (1993) and Widaman (1985), confirmatory factor analysis was used in the evaluation of multitrait-multimethod matrices. The measured variables included the nine *College BASE* cluster scores and the four cognitive-development scales from the senior survey. Because of significant multivariate skewness in the data, asymptotically distribution free estimation procedures were employed (Jöreskog and Sörbom, 1989).

Initially, four distinct factor structures were specified and test. The first model contained six latent variables, two representing the measurement methods and four representing the domains of English, mathematics, science, and social studies. The two latent variables representing the measurement methods were allowed to correlate freely, as were the four latent variables representing the outcomes domains. Method and trait factors were not correlated with each other. This model is depicted in Figure 1.

Insert Figure 1 about here

The second model in the multitrait-multimethod analysis contained the two freely correlated measurement factors, but the four outcomes factors were dropped from this model. A comparison of fit statistics for the first and second models represented a test of the extent to which the outcomes domains (traits) were needed to explain the relationship between measured variables. This comparison represented a direct test of convergent validity.

The third and fourth models contained the same six factors as the first model. In the third model the four latent variables representing outcomes domains were specified as being perfectly correlated. Comparison of the third and first models provided a test of whether self reports and test scores were able to discriminate among traits. Acceptance of the third model would indicate that the observed measures were not able to discriminate among traits. In the fourth model the two latent variables representing measurement methods were specified as being perfectly correlated. Comparison of this model to the first model provided a test of discrimination between measurement methods.

As a final step in the specification and estimation of models, a limited specification search was undertaken to identify the most appropriate model for representing the data.. This specification search involved testing whether it was possible to fix any of the correlations among latent variables to either 1.00 or 0.00 without significantly reducing goodness of fit.

Byrne (1993) suggested that models be compared using traditional chi-square goodness-of-fit measures and incremental fit indices. In this study, chi-square measures were

used, but incremental fit indices were not used because asymptotically distribution free estimation procedures tend to produce very poor estimates of model fit for a null model. Poor estimation of the null model produces inappropriate and unstable incremental fit indices for the higher-order models tested in a multitrait-multimethod analysis (Sugawara and MacCallum, 1993). Instead, a nonincremental fit index, Cudek and Browne's (1983) rescaled Akaike (1973) Information Criterion (RAIC) was used to compare model fit. This fit index has been shown to be robust with respect to departures from multivariate normality (Williams and Holahan, 1994). Smaller RAIC values indicate a better-fitting model.

As a final step in the multitrait-multimethod analysis, standardized factor loadings and squared multiple correlations for the measured variables were examined to determine if self reports and test scores measured the same constructs (Byrne, 1993). The analyses employed at this point in the study paralleled Jöreskog's (1971) procedures for th analysis of congeneric tests. In order to be considered measures of the same construct, trait factor loadings for the observed variables should be statistically significant and substantial.

In the second phase of the data analysis, a structural equation model of the relationships between environment, involvement, and interaction and *College BASE* scores was tested. Again, departures from multivariate normality required the use of asymptotically distribution free estimation procedures. The model included 23 measured variables and 10 latent variables. Six of the latent variables represented self reports of college experiences, while four latent variables represented outcomes domains. The model was specified so that the outcomes domains were "explained" by the college experience variables. Standardized effects coefficients and squared

multiple correlations for the structural equations were used to identify specific relationships and the strengths of those relationships.

Results

Self Reports as Proxies

Table 1 presents the goodness-of-fit results for the four initial models in the multitrait-multimethod analysis. As indicated in the table, the baseline model containing freely correlated method and trait factors produced a chi-square value of 127.09 ($df = 45$; $p < .001$). The rescaled AIC value for the model was 0.41. The second model, containing two methods factors but no trait factors, produced a much poorer fit to the data ($\chi^2 = 434.97$; $df = 64$; $p < .001$; RAIC = 0.91). A comparison of the first two models provides clear support for convergence. Eliminating the four trait factors significantly worsened model fit ($\Delta\chi^2 = 307.88$; $\Delta df = 19$; $p < .001$).

Insert Table 1 about here

The third model, containing methods factors and perfectly correlated traits factors, produced a chi-square value of 214.25 ($df = 51$; $p > .001$). When this model was compared to the baseline multitrait-multimethod model, it was found that adding the restriction that trait factors be perfectly correlated significantly decreased goodness of fit ($\Delta\chi^2 = 87.16$; $\Delta df = 6$; $p < .001$). Adding the restriction to the baseline model that methods factors be perfectly correlated (model 4), also significantly decreased goodness of fit relative to the baseline model

($\Delta\chi^2 = 162.04$; $\Delta df = 1$; $p < .001$). These results for the third and fourth models provide clear evidence of discrimination among the four outcomes traits and between the two measurement methods.

Although the baseline model represented a satisfactory explanation of the relationships among the observed measures, an examination of the parameter estimates in the baseline model revealed that the two methods factors were essentially uncorrelated ($\phi_{21} = 0.02$). Consequently, a fifth model, in which the correlation between the methods factors was set to zero, was specified and tested. Results indicated that adding this restriction to the model did not significantly decrease goodness of fit ($\Delta\chi^2 = 2.64$; $\Delta df = 1$; $p > .05$). Therefore, this model was used in subsequent evaluations of factor loadings for the multitrait-multimethod analysis.

Table 2 presents the factor loadings and squared multiple correlations for the final model. Also included (in parentheses) are estimates of the amount of variance in the observed measures that was explained by the method and trait factors. These estimates provide an indication of the relative explanatory power of the method and trait factors.

Insert Table 2 about here

An examination of the results in Table 2 reveals that the four survey scales all had significant positive loadings on survey-method factor. The lowest factor loading was 0.50 for the social-studies scale, while the highest loading was found for the science scale (0.80). Likewise, the nine *College BASE* cluster scores all had significant positive loadings on the test factor. Factor loadings ranged from a low of 0.41 for algebra to a high of 0.84 for the social science

cluster score. In contrast, the four scales from the senior survey had relatively weak loadings on the trait factors, ranging from a nonsignificant loading of 0.07 for the social studies survey scale to a loading of 0.37 for the mathematics survey scale. For all of the trait factors except mathematics, at least one of the *College BASE* cluster scores had very low loadings. This is evident in the proportion of variance in the cluster scores explained by the trait factors.

Based on an analysis of the factor loadings in the final model, it seems reasonable to conclude that evidence of convergent validity is strongest for the mathematics trait. Evidence of convergent validity for the English and science traits is relatively weak, and there is little evidence of convergent validity for the social-studies trait. Importantly, the survey items for social studies had the poorest content correspondence with the *College BASE* social studies cluster scores.

Self Reports as Policy Indicators

Table 3 presents the effects parameters for the relationships between latent variables representing college experiences and latent variables representing *College BASE* subject domains. Also included in the table are squared multiple correlations for the structural equations. These squared multiple correlations provided an indication of the explanatory power of the college experiences constructs. An examination of the effects parameters for college experiences and the English domain reveals that two of the self-report factors, writing and interaction with faculty and peers, were statistically significant (0.23 and 0.21, respectively). The squared multiple correlation for the structural equation representing the relationships between college experiences and English test scores was 0.15.

Insert Table 3 about here

The results in Table 3 indicate that college experiences were strongly related to mathematics test scores. Latent variables representing the college environment ($\gamma_{12} = 0.27$), academic involvement ($\gamma_{22} = 0.22$), and extracurricular involvement ($\gamma_{42} = 0.65$) all had significant positive effects on mathematics, while writing ($\gamma_{32} = -0.32$), cultural involvement ($\gamma_{52} = -0.37$) and interaction with faculty and peers ($\gamma_{62} = -0.45$) had significant negative effects on mathematics. The squared multiple correlation for the structural equation explaining mathematics outcomes was 0.25.

Similar results were obtained for the structural equation relating college experiences to the latent variable for science outcomes. The college environment ($\gamma_{13} = 0.27$), academic involvement ($\gamma_{23} = 0.43$), and extracurricular involvement ($\gamma_{53} = 0.52$) variables were positively and significantly related to science outcomes, while writing ($\gamma_{33} = -0.50$), cultural involvement ($\gamma_{53} = -0.53$), and interaction with faculty and peers ($\gamma_{63} = -0.35$) produced negative effects coefficients that were statistically significant. The squared multiple correlation for the structural equation explaining science outcomes was 0.28.

Of the six latent variables representing college experiences, only extracurricular involvement ($\gamma_{44} = 0.32$) was significantly related to social studies outcomes. The squared multiple correlation for the structural equation explaining social studies outcomes also was quite low (0.07).

Discussion

With respect to the two foci in this study, the results of the data analyses can be summarized as follows: First, the multitrait-multimethod analyses provided limited support for using self reports of cognitive development during college as proxies for *College BASE* test results. The general analysis of confirmatory factor analysis models representing convergence and discrimination indicated that self reports and test scores did converge with each other. The models also indicated that self reports and test scores were capable of discriminating among different dimensions of college outcomes. However, careful examination of the factor loadings in the final model revealed that there was strong evidence of convergence only for the mathematics domain. Evidence of convergence in the English and science domains was relatively weak, and no statistically significant evidence of convergence was found for the social studies domain. Based on these findings, it seems that the mathematics survey and *College BASE* mathematics scores measured the same construct, while the social studies survey scale and *College BASE* social-studies scores do not measure the same construct. Conclusions regarding the English and science domains is more ambiguous and should await further study.

With regard to the second research focus, reasonably strong evidence was found for using self reports of college experiences as policy indicators, particularly in the domains of mathematics and science. In both the mathematics and science domains, the squared multiple correlations for the structural equations indicated that the college experiences included in the study had a substantively meaningful relationship with test scores. The strength of the relationship between reported college experiences and English scores was somewhat weaker, nevertheless it was substantively meaningful. Consistent with the results in the first phase of the

research, self reports of college experiences were least strongly related to the social-studies domain.

Also of interest are the specific relationships between self reports of college experiences and educational outcomes. Consistent with theory, involvement in writing and interaction with faculty and peers were positively related to English performance. Somewhat surprising was the lack of significant effects for the college environment, academic involvement, and cultural involvement.

The effects of college experiences on math and science achievement produced mixed results. Consistent with expectations, perceptions of the college environment and academic involvement were positively related to math and science outcomes. Somewhat surprising was the strength of the relationship between extracurricular involvement and math and science achievement. While significant effects for involvement in writing and cultural involvement were not unexpected, the presence of a significant negative effect for experience in writing was counter to the results reported by Pascarella, Terenzini, and their colleagues. Perhaps most disturbing was the absence of significant relationships between college experiences and social-studies achievement. The absence of significant effects for social studies, plus the counterintuitive results for English, mathematics, and science, indicate that additional research is needed.

Although this study was limited to data from a single institution, the results do have implications for a national assessment of college student outcomes and for future research. Regarding implications for a national assessment, the present study provides a highly qualified "yes" to the question of whether self reports of cognitive development can be used as proxies for achievement-test results. Two qualifications are particularly important. First, this research

underscores the importance of having high content correspondence between self-report measures and test domains. In the present study, the convergence of self reports and test scores was greatest for the mathematics domain where content correspondence was relatively high. In the social studies domain, were content correspondence was relatively low, there was no statistically significant evidence of convergence. If assessment professionals are committed to developing self-report measures of cognitive development that can serve as proxies for test scores, the present research suggests that it may be necessary to use the same set of content specifications for self reports and test scores to provide evidence of strong convergent validity.

The second qualification to emerge from the current research is that comparing self reports and test scores is not a simple matter. The presence of a substantial amount of method-specific variance in self reports and test scores, coupled with the fact that the variance attributable to the two methods is orthogonal (unrelated), indicates that simple comparisons of self reports and test scores can lead to incorrect conclusions. The use of self reports as proxies for test scores may be possible, but it will require the use of sophisticated statistical techniques that can remove the method-specific biases from assessment data.

The answer to the question of whether self reports of college experiences can be used as policy indicators to improve student performance on achievement tests, is a much stronger "yes." The present study revealed that self reports of college experiences are significantly and meaningfully related to test scores at the University of Tennessee, Knoxville. Moreover, the generalizability of this finding seems reasonably strong, given that many of the observed relationships were generally consistent with the results of previous research. Additional research is needed, however, to determine if the counterintuitive results described above are anomalies.

Research is also needed to identify other aspects of students' college experiences that may be related to objective measures of college outcomes. The need for additional research is particularly great in the area of social-studies achievement where the college-experience measures used in the current study were only weakly related to achievement in this domain.

The clearest conclusion to emerge from this study is that before self reports can be used as either proxies or policy indicators, much more research is needed. Additional research is needed to identify those factors in students' college experiences that are most strongly related to learning outcomes and to clarify the effects of content correspondence on convergence and discrimination. Research is also needed to identify the best methods of representing the self reports of cognitive development absent the biasing effects of measurement methods. Most important, multi-institution studies should be conducted to determine if the relationships between self reports and test scores are stable across different types of institutions. As Ewell, Lovell, Dressler, and Jones (1993) pointed out, self reports can serve as valid proxies and/or policy indicators only if the relationships between self reports and objective tests of student achievement are consistent across institutions.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.), Second international symposium on information theory. Budapest: Akademiai Kiado.

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

Association of American Colleges (1985). Integrity in the college curriculum: A report to the academic community. Washington, D. C.: Author.

Astin, A. W. (1993). What matters in college? San Francisco: Jossey-Bass.

Baird, L. L. (1976a). Structuring the environment to improve outcomes. In O. Lenning (ed.), Improving educational outcomes (New Directions for Higher Education Series, no. 16, pp. 1-24). San Francisco: Jossey-Bass.

Baird, L. L. (1976b). 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 of 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.

Boyer, E. L. (1987). College: The undergraduate experience in America. New York: Harper & Row.

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.

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

Cudeck, R., and Browne, M. W. (1983). Cross-validation of covariance structures. Multivariate Behavioral Research, 18, 147-167.

Davis, T. M. and Murrell, P. H. (1990). Joint factor analysis of the College Student Experiences Questionnaire and the ACT COMP objective exam. Research in Higher Education, 31, 425-442.

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. ERIC Document Reproduction Service No. ED 340 755.

El-Khawas, E. (1990). Campus trends, 1990. Washington, D. C.: American Council on Education.

Elliott, E. (1991). Charge to participants. In A. Greenwood (ed.), National assessment of college student learning: Issues and concerns. Washington, D. C.: 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, pp. 75-125). Washington, D. C.: American Educational Research Association.

Ewell, P. T., Lovell, C. D., Dressler, P., and Jones, D. P. (1993). A preliminary study of the feasibility and utility for national policy of instructional "good practice" indicators in undergraduate education. Boulder, CO: National Center for Higher Education Management Systems.

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.

Friedlander, J. H. (1980). The importance of quality of effort in predicting college student attainment. Unpublished doctoral dissertation, University of California, Los Angeles.

Friedlander, J. H. (1991). The quality of students' educational experiences at Santa Barbara City College. Unpublished manuscript, Santa Barbara City College, Santa Barbara, CA.

Grossman, R. J. (1988). The great debate over institutional accountability. College Board Review, 147, 4-6, 38-42.

Hartle, T. W. (1986). The growing interest in measuring the educational achievement of college students. In C. Adelman (ed.), Assessment in American higher education: Issues and contexts (pp. 1-12). Washington, D. C.: U. S. Government Printing Office.

House, E. R. (1993). Professional evaluation: Social impact and political consequences. Newbury Park, CA: Sage.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests Psychometrika, 36, 109-133.

Jöreskog, K. G., and Sörbom, D. (1989). LISREL 7 user's reference guide. Chicago: Scientific Software.

National Education Goals Panel Resource Group on Adult Literacy and Lifelong Learning. (1991). Adult literacy and lifelong learning. In National Eduction Goals Panel, Measuring progress toward the national education goals: Potential indicators and measurement strategies (pp. 81-98). Washington, D. C.: U. S. Government Printing Office.

National Governors' Association. (1986). A time for results: The governors' 1991 report on education. Washington, D. C.: Author.

National Institute of Education Study Group on the Conditions of Excellence in American Higher Education (1984). Involvement in learning: Realizing the potential of American higher education. Washington, D. C.: U. S. Government Printing Office.

Osterlind, S. J. (1989). College BASE: Guide to test content. Chicago, IL: Riverside.

Pace, C. R. (1987). CSEQ test manual and norms. Los Angeles: Center for the Study of Evaluation.

Pace, C. R. (1990). The undergraduates: A report on their activities and progress in college in the 1980s. Los Angeles: Center for the Study of Evaluation.

Pascarella, E. T., and Terenzini, P. T. (1978). Student-faculty informal relationships and freshman year educational outcomes. Journal of Educational Research, 71, 183-189.

Pike, G. R. (1992). A generalizability analysis of the College Basic Academic Subjects Examination. Unpublished manuscript, Center for Assessment Research and Development, Knoxville, TN.

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.

Porter, O. T. (1982). The role of quality of effort in defining institutional environments: An attempt to understand college uniqueness. Unpublished doctoral dissertation, University of California, Los Angeles.

Ratcliff, J. L. (1991). What type of national assessment fits American higher education. 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, H. M., and MacCallum, R. C. (1993). Effect of estimation method on incremental fit indexes for covariance structure models. Applied Psychological Measurement, 17, 365-378.

Terenzini, P. T., Pascarella, E. T., and Lorang, W. (1982). An assessment of the academic and social influences on freshman year educational outcomes. Review of Higher Education, 5, 86-110.

Terenzini, P. T., and Wright T. (1987). Influences on student's academic growth during four years of college. Research in Higher Education, 26, 161-179.

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.

Table 1:

Goodness-of-fit Results for the Models Used in the Multitrait-Multimethod Analysis

Model	df	χ^2	RAIC	Δdf	$\Delta\chi^2$
(1) Baseline	45	127.09[a]	0.41	---	---
(2) No Traits	64	434.97[a]	0.91	19	307.88[a]
(3) Perfectly Correlated Traits	51	214.25[a]	0.55	6	87.16[a]
(4) Prefectly Correlated Methods	46	289.13[a]	0.71	1	162.04[a]

[a] $p < .001$

Table 2:

Standardized Measurement Parameters, Variance Estimates, and Squared Multiple Correlations for Observed Variables in the Final Model

Observed Variables	Latent Variables						SMC
	Survey	Test	English	Math	Science	S. Studies	
English Reports	0.60[c]		0.17[c]				0.39
	(0.36)		(0.03)				
Math Reports	0.64[c]			0.37[c]			0.54
	(0.41)			(0.14)			
Science Reports	0.80[c]				0.12[b]		0.66
	(0.64)				(0.01)		
Social Studies	0.50[c]					0.07	0.26
Reports	(0.25)					(0.01)	
CB Reading and		0.71[c]	0.26[c]				0.58
Literature		(0.50)	(0.07)				
CB Writing		0.52[c]	0.66[c]				0.71
		(0.27)	(0.44)				

Observed Variables	Latent Variables						SMC
	Survey	Test	English	Math	Science	S. Studies	
CB General Math		0.58[c]		0.62[c]			0.72
		(0.34)		(0.38)			
CB Algebra		0.41[c]		0.72[c]			0.70
		(0.17)		(0.52)			
CB Geometry		0.53[c]		0.64[c]			0.68
		(0.28)		(0.41)			
CB Lab and Field		0.77[c]			0.61[c]		0.96
Work		(0.59)			(0.37)		
CB Fundamental		0.72[c]			0.21[b]		0.56
Concepts		(0.52)			(0.04)		
CB History		0.79[c]				0.50[c]	0.87
		(0.62)				(0.25)	
CB Social Science		0.84[c]				0.14[b]	0.72
						(0.02)	

[a]$p < .05$; [b]$p < .01$; [c]$p < .001$

Table 3:

Standardized Effects Parameters and Squared Multiple Correlations for Relationships Between Reports of College Experiences and Test Scores

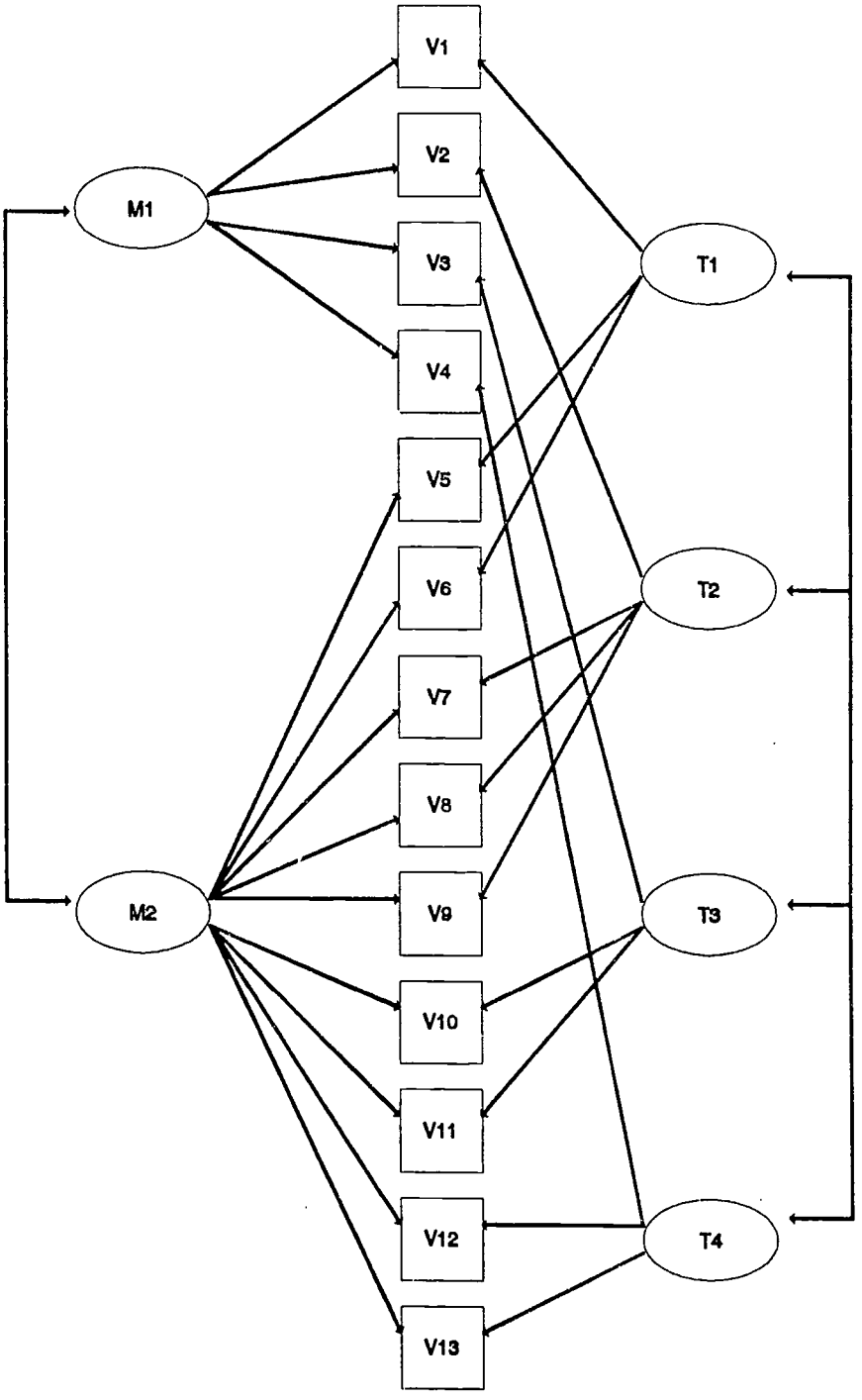
	English	Math	Science	S. Studies
Environment	0.03	0.27[a]	0.27[a]	0.05
Academic Involvement	0.08	0.22[a]	0.43[b]	0.00
Writing	0.23[a]	-0.32[b]	-0.50[b]	0.02
Extracurricular Involvement	0.01	0.65[b]	0.52[b]	0.32[a]
Cultural Involvement	0.04	-0.37[b]	-0.53[b]	-0.13
Interaction with Faculty/Peers	0.21[a]	-0.45[b]	-0.35[b]	-0.05
SMCs for the Structural Equations	0.15	0.25	0.28	0.07

[a] $p < .05$; [b] $p < .01$

Figure 1:

A Simplified* Baseline Model for the Multitrait-Multimethod Analysis

*Note: Measurement errors have been omitted to improve readability.


V1
V2
M1
V3
T1
V4
V5
V6
T2
V7
V8
M2
V9
T3
V10
V11
V12
T4
V13