The Convergent and Discriminant Validity of NSSE Scalelet Scores

Gary R. Pike

Faculty and administrators are more likely to take responsibility for student learning and development if they believe that assessment data represent their students and identify specific actions for improvement. An earlier study found that NSSE scalelets provide dependable metrics for assessing student engagement at the university, college, and department levels. Building on the earlier study, the findings of the current research indicate that the NSSE scalelets have greater explanatory power and provide richer detail than the NSSE benchmarks.

Assessment has become an integral part of American higher education, and surveys of constituents are an important element in assessment efforts. The National Center for Postsecondary Improvement (NCPI) reported that 96% of the 1400 institutions responding to its survey had implemented some form of assessment, and 75% used surveys in their assessment efforts (Peterson, Einarson, Augustine, & Vaughan, 1999). Although assessment is widespread, examples of assessment data, including survey results, being used to effect institutional change are relatively rare (Banta, 2002; Ewell, 2002; Peterson et al.). Pike (2002, p. 147) concluded "there is no greater problem in assessment than our inability to influence decision making with assessment results."

A major barrier to using survey data for improvement is that many campus decision makers, particularly deans and department heads, find the results of institutional surveys to be too general (Kuh, Gonyea, & Rodriguez, 2002). That is, the results do not suggest specific courses of action. Experience indicates that faculty and administrators are more likely to take responsibility for student learning and development if they believe that assessment data represent their students and identify specific actions for improvement. Case studies of effective use of survey results reveal that surveys lead to improvement when the data are broken down or disaggregated at the college or department level and focus on a few highly related items that suggest specific actions (El-Khawas, 2003; Kezar, 2002, 2003). Presenting results that are specific to a department or college frequently requires that a survey be administered to a large number of students in order to produce dependable measures (Indiana University Center for Postsecondary Research, 2001; Kuh, Gonyea, & Rodriguez, 2002). This is a requirement that institutions may not be able to meet for many of their programs.

In an earlier article, I proposed that researchers and assessment professionals use scalelets to overcome the challenges posed by the need to present survey data that are specific to a department or college (Pike, 2006). My generalizability study found that the 12 NSSE scalelets I developed yielded dependable college experience mean scores based on relatively few (i.e., 25 to 50) respondents. The present research builds on my earlier generalizability study and examines the convergent and discriminant validity of using the 12 NSSE scalelet scores in assessment research.

Gary R. Pike is the Director of Institutional Research at Mississippi State University.

BACKGROUND

The term scalelet is derived from the concept of testlets proposed by Wainer and Kiely (1987, p. 190): "A testlet is a group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee may follow [in a computerized adaptive test]." The use of testlets allows developers to construct test units that contain more than one item and reduces problems associated with context and order effects (Wainer et al., 1990). The result is test scores with greater dependability and less error than scores based on single items (Thissen & Mislevy, 1990).

A scalelet consists of a set of survey questions related to a specific aspect of the educational experiences of a group of students. Three elements of this definition require elaboration. First, a scalelet consists of a set of survey questions. It is not possible to make generalizations about a construct, such as involvement in cocurricular activities, based on a single survey question (e.g., the number of organizations to which a student belongs). The richness of the constructs used in outcomes assessment requires that generalizations based on a single question be limited to that question. In the preceding example, generalizations should be made about the number of organizations to which the student belongs.

Although the ideal would be to base generalizations about students' experiences on all possible questions about those experiences, the reality of survey research requires that assessment professionals base their conclusions on responses to a sample of questions. The second element in the definition of scalelets—that questions relate to a specific educational experience—allows a relatively small sample of items to be included in a scalelet. In effect, there is a continuum ranging from very broad generalizations, based on many survey questions, to specific conclusions based on a single item. Scalelets strike a balance between the breadth of generalizations and the number of questions in a survey.

The third element—that scalelets represent the experiences of groups of students—is based on an understanding that assessment for accountability or program improvement requires data about groups (Ewell, 1991; Pike, 1994). The cocurricular experiences of a student are important, but data about a single student provides little information about programs at the institution. Evaluations of the cocurriculum should ideally be based on data about the experiences of all students or at least a representative sample of students.

The use of scalelets requires that researchers and assessment professionals make several different generalizations from samples to populations. Evaluating the quality and effectiveness of a program requires that an assessor make generalizations about the effectiveness of a program based on a sample of questions about the program. Likewise, the assessor may need to make generalizations about all of the students in a program based on a sample of students in that program. Generalizability theory, developed by Cronbach, Gleser, Nanda, and Rajaratnam (1972), provides a mechanism for researchers, assessment professionals, and policy makers to identify the limits of the inferences they can draw from their samples (Brennan, 1983; Shavelson & Webb, 1991).

In an earlier study, I examined the dependability (i.e., generalizability) of 12 scalelets drawn from questions used by the National Survey of Student Engagement (NSSE). Because the scalelets required that generalizations be made over samples of items and samples of students, the generalizability of group means was the focus of the study (Kane, Gillmore, & Crooks, 1976; Pike, 1994). Based on the responses of 50 randomly selected

seniors each from 50 randomly selected institutions, I found that all 12 scalelets produced dependable group means ($E\rho^2 \geq 0.70$) with 25 to 50 respondents (Pike, 2006).

Although dependability is a necessary condition for demonstrating that scalelets can provide valid scores for assessment research, it is not sufficient (Messick, 1989). Additional criteria must be satisfied. Banta and Pike (1989) have argued that these criteria should include the convergent and discriminant validity of scores.

Given that the objective of using scalelets in outcomes assessment is to make judgments about educational quality, validity provides a reference point for evaluating scalelets because questions about validity focus on the adequacy and appropriateness of the inferences drawn from data (Cronbach, 1971; Messick, 1989). Although several criteria can be used to evaluate the validity of a measure, there is a growing sentiment that validity should be treated as a unitary concept with construct validity at its core (Angoff, 1988; Messick). Loevinger (1957) grouped questions related to construct validity into three categories: (a) the extent to which items are accounted for by the construct (a substantive component), (b) the extent to which relationships among the items reflect relationships within the construct (a structural component), and (c) the extent to which relationships between scores and other variables are consistent with theories of the construct (an external component).

The approach used in this study is based on Loevinger's (1957) external component of construct validity and focuses on the concepts of convergence and discrimination (Campbell, 1960; Campbell & Fiske, 1959; Fiske, 1982). Banta and Pike (1989) and Pike (1989, 1992) used this approach to evaluate several tests of general-education outcomes. For an assessment measure to be a valid indicator of program effectiveness, scores should be associated (i.e., converge) with other measures of educational quality and student learning. Moreover, these relationships should transcend institutional characteristics such as size, selectivity, mission, and control. In addition, the scores should discriminate among different quality indicators. That is, scores for some measures should be associated with one set of learning outcomes, whereas the scores for other measures should be associated with different learning outcomes. Absence of discrimination would indicate that scalelets are not needed and that a total score would be a sufficient indicator of program quality.

Because the purpose of this research was to evaluate measures of student engagement, student-engagement theory served as the construct against which the scalelets were judged. This theory has its origin in the work of Pace (1980, 1984), Astin (1984, 1985), and Kuh and his colleagues (Kuh, Schuh, Whitt, & Associates, 1991). Although the writers used different terminology (e.g., quality of effort, involvement, and engagement) to describe their concepts, their views were based on the deceptively simple premise that students learn from what they do (Kuh, 2003). A second important premise of student engagement theory is that, even though the focus is on *student* engagement, *institutional* actions influence levels of engagement and learning on campus (Astin, 1985; Kuh, Schuh, et al.; Pace, 1984).

Research has provided consistent support for both assumptions. Studies show that engagement is positively related to test scores and students' reports of learning (Gellin, 2003; Kuh, Hu, & Vesper, 2000; Pascarella et al., 1996; Pike, 1995; Pike, Kuh, & Gonyea, 2003). Moreover, different types of engagement have been found to be differentially related to learning outcomes. For example, Pike (1995) found that students' writing experiences and their interactions with faculty

and peers were positively related to English outcomes but negatively related to learning in mathematics. Conversely, involvement in academic activities and extracurricular involvement were not related to learning in English but were positively related to learning in mathematics and the social sciences.

Within the context of student-engagement theory, it is reasonable to expect that measures of student engagement will converge with and discriminate among measures of student learning. Two questions, corresponding to the concepts of convergence and discrimination, formed the basis for the current research:

1. Are NSSE scalelet scores significantly related to institutional measures of student learning and development after accounting for institutional characteristics?
2. Are NSSE scalelet scores differentially related to student learning outcomes after accounting for institutional characteristics?

The answer to the first question provides evidence of convergent validity, whereas the answer to the second question provides evidence of discriminant validity.

RESEARCH METHODS

Data Sources

The data for this study came from the 2004 administration of the NSSE survey, the Integrated Postsecondary Education Data System (IPEDS) data files, Barron's ratings of institutional selectivity, and institutional enrollment reports. The initial sample consisted of 114,061 seniors attending 473 four-year institutions. A comparison of the characteristics of the NSSE 2004 institutions and all four-year colleges and universities revealed that NSSE institutions were very similar to the national profile in terms of geographic region and urban-rural location. Public institutions and master's universities were over-represented among survey participants, whereas baccalaureate general colleges were under-represented (Indiana University Center for Postsecondary Research, 2004).

At the conclusion of the survey cycle, 45,208 seniors had responded to the survey, a response rate of slightly less than 40%. A comparison of respondents' characteristics to the characteristics of the student populations at participating institutions revealed that women were over-represented among respondents, as were Caucasians and full-time students. However, the differences were relatively small and should not affect the generalizability of the results (Indiana University Center for Postsecondary Research, 2004). Institutional means based on seniors' responses were used in this study. Complete data, including institutional benchmark and scalelet scores based on the responses of at least 50 seniors, were available for 454 colleges and universities.

Measures

Forty-nine questions from the NSSE survey were used to create 12 scalelets. A list of the questions comprising the scalelets along with generalizability coefficients for group means based on 50 students are included in the appendix. The items comprising each scalelet were selected based on face and content validity, and the content of the scalelets paralleled the content of the NSSE benchmarks. For example, most of the items in the Course Challenge, Writing Experiences, and Higher-Order Thinking Skills scalelets were drawn from the Level of Academic Challenge benchmark. Items included in the Active Learning and Collaborative Learning scalelets were from the Active and Collaborative Learning benchmark, and items in the Course Interaction and Out-of-Class Interaction

scalelets were from the Student Interaction with Faculty Members benchmark. Many of the items in the Varied Experiences and Information Technology scalelets were taken from the Enriching Educational Experiences benchmark. The Diversity Experiences scalelet also was composed of items from the Enriching Educational Experiences benchmark. The items included in the Support for Student Success and Interpersonal Environment scalelets came from the Supportive Campus Environment benchmark.

Questions about gains in learning were used to create two outcome measures. Originally developed by Kuh, Gonyea, and Palmer (2001), the Gains in General Education scale includes questions about gains in writing, speaking, analytical skills and general education. The Gains in Practical Competence scale includes gains in computer and information technology, quantitative skills, and knowledge and skills needed for work. In addition to the scalelets and gain scores, the NSSE benchmarks were included in the study, and the items comprising the benchmarks are identified in the appendix.

Several institutional characteristics were included in the study. These variables were institutional control (1 = Private, 0 = Public), Carnegie classification (dummy coded as Doctoral/Research-Extensive, Doctoral/Research-Intensive, Master's, Baccalaureate Liberal Arts, and Baccalaureate General [not coded]), percent of female students, percent of minority students, percent of on-campus students, and percent of full-time students. These measures were taken from IPEDS data. Two other characteristics, Barron's selectivity ratings and Fall 2003 enrollment as reported by the institutions, were included in the study.

Data Analysis

Institutions served as the units of analysis in the study. Initially, correlations among institutional characteristics, NSSE benchmarks, scalelets, and outcome measures were calculated to aid in interpreting the results of subsequent analyses. Next, four multiple regression models were specified and tested. In the first model, general-education gains were regressed on institutional characteristics and NSSE benchmarks. In the second model, general-education gains were regressed on institutional characteristics and scalelet scores. Gains in practical skills were regressed on institutional characteristics and NSSE benchmarks in the third model, and practical skill gains were regressed on institutional characteristics and scalelet scores in the final model.

Testing of regression models was a two-step process. First, gain scores were regressed on institutional characteristics. Second benchmark or scalelet scores were added to the models. Significance tests and measures of explained variance were calculated for both steps to evaluate convergence. Significance tests provided indications of whether there were relationships between benchmarks or scalelets and gains, whereas changes in explained variance indicated whether the relationships were educationally important. Standardized regression coefficients from the second step identified unique relationships between benchmarks or scalelets and gains. Different patterns of relationships across scalelets and gains indicated that the scalelets were able to discriminate among different types of engagement and different learning outcomes.

Tests of multicolinearity and influence diagnostics were calculated for each model. Preliminary analyses indicated that multicolinearity was not a problem. However, examination of the influence statistics revealed that two institutions had extreme scores that exerted undue influence on the regression results. Both institutions were small private

liberal arts colleges with extremely high levels of student engagement. Those institutions were dropped from the final analyses.

RESULTS

Table 1 displays the correlations and standardized regression coefficients representing the relationships between the gain measures and institutional characteristics, NSSE benchmarks, and scalelet scores. Most of the independent variables included in the regression analyses were significantly correlated with general education gains, and many of the independent variables were significantly correlated with gains in practical skills. Moreover, the results of the regression analyses provided clear evidence of the convergent validity of the NSSE benchmarks and scalelet scores. Institutional characteristics and NSSE benchmarks accounted for 78.0% of the variance in general-education gains, and the NSSE benchmarks alone accounted for 30.7% of the variance. The relationships were slightly stronger for the model that included scalelet scores. Institutional characteristics and scalelet sores accounted for 81.3% of the variance in general-education gains and the scalelets accounted for 34.0% of the gain-score variance.

Evidence supporting the convergent validity of scalelet scores was more pronounced for gains in practical skills. Institutional characteristics and NSSE benchmarks combined to explain 40.3% of the variance in practical-skill gains, and 22.2% of this variance was explained by the benchmarks. Institutional characteristics and scalelet scores explained 53.6% of the variance in gains in practical skills. Of the variance in practical-skill gains, 35.5% was explained by the scalelet scores.

The standardized regression coefficients, shown in Table 2, also provide evidence of convergent validity. However, caution should be used in interpreting the coefficients due to the intercorrelations among institutional characteristics and engagement measures. Statistically significant regression coefficients that have the same sign as the corresponding correlations, which are also statistically significant, indicate that the variables uniquely contribute to the variance in a gain score. Nonsignificant coefficients or coefficients with signs that are opposite the signs of the corresponding correlations indicate that the variable does not uniquely contribute to the gain measure.

Evidence of the convergent validity of the NSSE scalelets can be found in the fact that gains in general education, which include gains in writing and analytical skills, were related to the Writing Experiences and Higher-Order Thinking Skill scores. Practical-skill gains, which include gains in understanding and using information technology, were positively related to scores for the Information Technology scalelet.

The multiple regression results also provide evidence of discriminant validity. Two types of evidence were found. First, the relationships between scalelet scores and gains were more highly differentiated than the relationships between gain scores and the NSSE benchmarks. For example, both the Course Interaction and Varied Experiences scalelets uniquely contributed to the variance in general-education gains, but the Student Interaction with Faculty Members and Enriching Educational Experiences benchmarks were not uniquely related to gains. Although the Active and Collaborative Learning benchmark was positively related to gains in practical skills, only the Collaborative Learning benchmark had a statistically significant relationship with this type of gain. The Enriching Educational Experiences benchmark was negatively related to practical-skill gains, but the relationships for the scalelets derived from this

TABLE 1.

Relations and Standardized Regression Coefficients for Gains in Student Learning

	General Education			Practical Skills		
	r	β_1	β_2	*r*	β_1	β_2
Private Control	0.511*	0.054	0.053	0.002	0.082	0.154*
Doc/Res-Extensive	-0.273*	-0.066	-0.044	0.006	0.183*	0.139*
Doc/Res-Intensive	-0.218*	-0.050	-0.043	0.071	0.143*	0.063
Master's	-0.133*	0.008	0.009	0.097*	0.053	0.026
Bac-Liberal Arts	0.499*	0.126*	0.154*	-0.310*	-0.230*	-0.172*
Bac-General	0.060			0.125*		
Selectivity	0.286*	0.072	0.091*	-0.172*	0.036	-0.062
Fall 2003 Enrollment	-0.443*	0.079	0.048	0.023	0.081	0.006
Percent Female	0.200*	0.015	0.002	-0.030	-0.199*	-0.182*
Percent Minority	-0.022	0.015	0.014	0.263*	0.099*	0.127*
Percent On-Campus	0.351*	-0.157*	-0.098*	-0.265*	-0.202*	-0.232*
Percent Full-Time	0.233*	-0.060	-0.035	-0.185*	-0.084	-0.052
Level of Acad. Challenge	0.796*	0.471*		0.017	0.081	
Active & Collab. Learning	0.561*	-0.011		0.222*	0.247*	
Interaction with Faculty	0.662*	0.099		-0.010	0.332*	
Enriching Educ. Exper.	0.624*	0.068		-0.184*	-0.463*	
Supportive Campus Envir.	0.706*	0.377*		0.215	0.372*	
Course Challenge	0.670*		0.165*	0.052		0.056
Writing Experiences	0.639*		0.182*	0.035		-0.124*
Higher-Order Thinking	0.660*		0.230*	0.262*		0.207*
Active Learning	0.563*		-0.007	0.164*		0.111
Collaborative Learning	0.408*		-0.022	0.230*		0.153*
Course Interaction	0.719*		0.196*	0.080		-0.065
Out-of-Class Interaction	0.571*		-0.231	-0.067		0.293*
Varied Experiences	0.594*		0.213*	-0.237*		-0.521*
Information Technology	0.268*		0.010	0.203*		0.318*
Diversity Experiences	0.404*		-0.048	-0.002		-0.176*
Support for Success	0.714*		0.235*	0.176*		0.217*
Interpersonal Environment	0.570*		0.131*	0.233*		0.193*
R^2		0.780	0.813		0.403	0.536
R^2 Change		0.307	0.340		0.222	0.355

*$p < 0.05$.

benchmark were quite different. Information Technology scores were positively related to gains, whereas Varied Experiences scores were negatively related to gains.

A comparison of the relationships between scalelet scores and gains across the two outcome measures provided additional evidence of discriminant validity. All three of the scalelets derived from the Level of Academic Challenge benchmark were related to general-education gains, but only the Higher-Order Thinking Skills scalelet was related to practical-skill gains. Neither the Active Learning nor the Collaborative Learning scalelets were related to general-education outcomes, but Collaborative Learning scores were related to gains in practical skills. Conversely, Course Interaction scores were related to general education but not practical-skill gains. Varied Experiences scores were positively related to gains in general education but negatively related to gains in practical skills. Information Technology scores were positively related to gains in practical skills.

Limitations

Although the results for NSSE 2004 are generally consistent with the results reported across the first few years of surveys, only one year of data was analyzed in this study. If institutions participating in other years were included, the results might differ in unknown ways. In addition, the data in this study are specific to the NSSE survey. Consequently, the results of this study do not indicate that valid scalelets can be developed for other surveys. Furthermore, the data for the validity analyses were at the institution, rather than the college or department level. If department-level data had been used, different results might have been obtained.

The most serious limitation is that the criterion variables for establishing convergent and discriminant validity were students' self-reports of their learning. Although self-report data have been studied extensively and shown to yield valid assessment information (see Kuh, 2001), both the measures to be evaluated and the criteria for evaluation used the same measurement method. Messick (1989) noted that the use of a single measurement method in validity studies may produce misleading results due to shared, method-specific variance. The presence of method-specific variance in this study may explain why most of the correlations between outcome measures, NSSE benchmarks, and scalelet scores were positive and statistically significant.

DISCUSSION

Despite these limitations, the results of the present research have important implications for assessment practice. For institutions that participate in the NSSE, this study indicates that NSSE scalelet scores provide valid measures of students' educational experiences and can be used for institutional assessment and improvement. The presence of strong relationships between scalelet scores and self-reported gains in student learning is one indication of the convergent validity of these scores. Scalelet scores accounted for approximately 35% of the variance in both gain measures. Scalelet scores also evidenced greater explanatory power than the NSSE benchmark scores. Increases in explained variance ranged from 3% for general-education gains to 13% for gains in practical skills. Most important, the relationships supporting the convergent validity of scalelet scores were consistent with student-engagement theory. That is, a particular type of involvement was associated with gains in a corresponding content area or skill. For example, greater involvement with writing was positively related to gains in general education, which included gains in writing. Likewise, experience with information tech-

nology was positively related to gains in practical skills, including gains in the ability to use information technology effectively.

The results of this study also provide evidence of the discriminant validity of NSSE scalelet scores. Generally, the relationships between engagement and outcomes were more nuanced for scalelet scores than for the NSSE benchmark scores. For example, the Active and Collaborative Learning benchmark was positively related to gains in practical skills. However, the analysis of scalelet scores revealed that this relationship was present for collaborative learning but not active learning. Similarly, the regression analyses indicated that scores on the Student Interaction with Faculty Members benchmark were positively related to gains in practical skills. Regression results for the scalelet scores indicated that out-of-class interaction with faculty members was related to gains in practical skills but interaction during class was not.

Once again, the evidence supporting the discriminant validity of scalelet scores is consistent with student-engagement theory. Take for example the relationship between the Varied Experiences and Information Technology scores and the two learning-outcome measures. Both the Varied Experiences and Information Technology scalelets are subsumed by the Enriching Educational Experiences benchmark. In the current study, Enriching Educational Experiences scores are not significantly related to gains in general education, but they are negatively related to gains in practical skills. Both results are surprising given that student-engagement theory suggests that many of the activities included in the benchmark, such as interacting with diverse groups of students, attending campus events, and participating in learning communities, are associated with gains in general education. Similarly, use of electronic technology, which is also included in the Enriching Educational Experiences benchmark, should be positively related to gains in practical skills. When gain scores were regressed on the scalelet scores, the results conformed to expectations. Scores on the Varied Experiences scalelet were positively related to gains in general education but negatively related to gains in practical skills. Scores on the Information Technology scalelet were positively related to gains in practical skills but not significantly related to gains in general education.

These results do not warrant abandoning the NSSE benchmark scores. Those scores serve a very useful purpose of providing senior administrators with a general overview of engagement on their campuses. Scalelet scores are most useful to academic affairs, student affairs, and assessment professionals who are charged with taking NSSE results and translating them into a series of action items to improve the student experience on campus. In addition, the present research should be considered a starting point for further research on the NSSE scalelets. The results of this study indicate that the NSSE scalelet scores can provide useful information for improvement at the institution level. More research is needed to demonstrate the convergent and discriminant validity of scalelet scores at the college and department levels. As Messick (1989) noted, validation is an on-going process of collecting and synthesizing information about the accuracy and appropriateness of scores for a variety of uses and across a variety of contexts.

The results of this research also have important implications for institutions that are involved in survey research but do not participate in NSSE. Scalelets are not unique to NSSE. Institutions that use other commercially available surveys, such as the College Student Experiences Questionnaire or the Cooperative Institutional Research Program freshman survey, may want to consider

developing scalelets that are specific to those instruments. The key to developing scalelets is in identifying a limited number of survey questions that are related to a specific aspect of students' educational experiences. Once possible scalelets have been identified, research can be conducted to evaluate the generalizability and validity of scalelet scores.

Scalelets can also be constructed for institutions using locally developed surveys. In fact, scalelets may hold the greatest promise for surveys developed by colleges and universities because they suggest a new model for survey construction. Wainer and Kiely (1987) argued that testlets are the basic building blocks of computer adaptive tests. Test questions are important only insofar as they contribute to the development of testlets. In an earlier article, I argued that the same model can be applied to survey development (Pike, 2006). The development process would begin with the identification of the constructs to be assessed by the survey. The definitions of these constructs would serve as the frameworks around which the scalelets would be developed. Samples of items would be generated and tested for each construct in order to identify groups of four or five questions that would yield generalizable and valid scalelet scores. The final step in the process would be to combine the items comprising the scalelets with biographic and demographic questions to form a completed survey.

Although identifying strategies for improving student learning was not an objective of this study, the results do suggest that some types of engagement initiatives may result in broad learning gains, whereas other engagement initiatives may yield more focused improvements. If an institution is interested in increasing gains in a broad range of learning outcomes, the evidence suggests that the institution should focus on strategies that would improve support for student success, the quality of the interpersonal environment, and students' higher-order thinking skills. Institutions interested in improving specific learning outcomes should focus on improving collaborative learning, course interaction, the variety of experiences available to students, and the use of information technology.

Correspondence concerning this article should be addressed to Gary R. Pike, Office of Institutional Research, Mississippi State University, 269A Allen Hall, Mississippi State, MS 39762-5708; gpike@ir.msstate.edu

Should the Appendix have a title?

APPENDIX.

Level of Academic Challenge Benchmark ($E\rho^2 = 0.71$)[a]

Course Challenge ($E\rho^2 = 0.73$)

- How often have you . . . worked harder than you thought you could to meet an instructor's standards or expectations? [workhard][b]
- How often have you . . . come to class without completing readings or assignments? {Reverse Scored} [clunprep]
- To what extent have . . . your examinations during the current school year challenged you to do your best work? [exams]
- How many hours a week do you spend . . . preparing for class (studying, reading, writing, rehearsing, and other activities related to you academic program)? [acadpr01][b]
- To what extent does your institution emphasize . . . spending significant amounts of time studying and on academic work? [envschol][b]

Writing ($E\rho^2 = 0.75$)

- How often have you . . . prepared two or more drafts of a paper or assignment before turning it in? [rewropap]
- How often have you . . . worked on a paper or project that required integrating ideas or information from various sources? [integrat]
- During the current school year . . . number of written papers or reports of 20 pages or more? [writemor][b]
- During the current school year . . . number of written papers or reports between 5 and 19 pages? [writemid][b]
- During the current school year . . . number of written papers or reports of fewer than 5 pages? [writesml][b]

Higher-Order Thinking Skills ($E\rho^2 = 0.77$)

- During the current school year, to what extent has your coursework emphasized . . . memorizing facts, ideas, or methods from your courses and readings so you can repeat them in pretty much the same form? {Reverse Scored} [memorize]
- During the current school year, to what extent has your coursework emphasized . . . analyzing the basic elements of an idea, experience, or theory, such as examining a particular case or situation in depth and considering its components? [analyze][b]
- During the current school year, to what extent has your coursework emphasized . . . synthesizing and organizing ideas, information, or experiences into new, more complex interpretations and relationships? [synthesz][b]
- During the current school year, to what extent has your coursework emphasized . . . making judgments about the value of information, arguments, or methods such as examining how others gathered and interpreted data and assessing the soundness of their conclusions? [evaluate][b]
- During the current school year, to what extent has your coursework emphasized . . . Applying theories or concepts to practical problems or in new situations? [applying][b]

Active and Collaborative Learning benchmark ($E\rho^2 = 0.81$)

Active Learning ($E\rho^2 = 0.84$)

- How often have you . . . asked questions in class or contributed to class discussions? [clquest][b]
- How often have you . . . made a class presentation? [clpresen][b]
- How often have you . . . participated in a community-based project as part of a regular course? [commproj][b]

Collaborative Learning ($E\rho^2 = 0.72$)

- How often have you . . . worked with other students on projects during class? [classgrp][b]
- How often have you . . . worked with classmates outside of class to prepare class assignments? [occgrp][b]
- How often have you . . . tutored or taught other students (paid or voluntary)? [tutor][b]
- How often have you . . . discussed ideas from your readings or classes with others outside of class (students, family members, coworkers, etc.)? [oocideas][b]

Student Interaction with Faculty Members benchmark ($E\rho^2 = 0.85$)

Course Interaction ($E\rho^2 = 0.80$)

- How often have you . . . discussed grades or assignments with an instructor? [facgrade][b]

continued

APPENDIX. *continued*

- How often have you . . . discussed ideas from your readings or classes with faculty members outside of class? [facideas][b]
- How often have you . . . received prompt feedback from faculty on your academic performance (written or oral)? [facfeed][b]

Out-of-Class Interaction ($E\rho^2 = 0.84$)

- How often have you . . . talked about career plans with a faculty member or advisor? [facplans][b]
- How often have you . . . worked with faculty members on activities other than coursework (committees, orientation, student-life activities, etc.)? [facother][b]
- Have you, or do you plan to, . . . work on a research project with a faculty member outside of course or program requirements? [research][b]

Enriching Educational Experiences Benchmark ($E\rho^2 = 0.79$)

Varied Experiences ($E\rho^2 = 0.94$)

- Have you, or do you plan to, . . . participate in a practicum, internship, field experiences, co-op experience, or clinical assignment? [intern][b]
- Have you, or do you plan to, . . . participate in community service or volunteer work? [volunteer][b]
- Have you, or do you plan to, . . . participate in a learning community or some other formal program where groups of students take two or more classes together? [learncom][b]
- Have you, or do you plan to, . . . take foreign-language coursework? [forlang][b]
- Have you, or do you plan to, . . . study abroad? [studyabr][b]
- Have you, or do you plan to, . . . participate in an independent study or self-designed major? [indstudy][b]
- Have you, or do you plan to, . . . participate in a culminating senior experiences (comprehensive exam, capstone course, thesis, project, etc.)? [seniorx][b]
- How many hours a week do you spend . . . participating in co-curricular activities (organizations, campus publications, student government, social fraternity or sorority, intercollegiate or intramural sports, etc.)? [cocurr01][b]
- To what extent does your institution emphasize . . . attending campus events and activities (special speakers, cultural performances, athletic events, etc.)? [envevent]

Information Technology ($E\rho^2 = 0.81$)

- How often have you . . . used an electronic medium (list-serv, chat group, Internet, etc.) to discuss or complete and assignment? [itacadem][b]
- How often have you . . . used e-mail to communicate with an instructor? [email]
- To what extent does your institution emphasize . . . using computers in academic work? [envcompt]

Diversity ($E\rho^2 = 0.77$)

- How often have you . . . had serious conversations with students of a different race or ethnicity than your own? [divrstud][b]
- How often have you . . . had serious conversations with students who differ from you in terms of their religious beliefs, political opinions, or personal values? [diffstu2][b]
- To what extent does your institution emphasize . . . encouraging contact among students from different economic, social, and racial or ethnic backgrounds? [envdivrs][b]

Supportive Campus Environment benchmark ($E\rho^2 = 0.84$)

Support for Student Success ($E\rho^2 = 0.83$)

- To what extent does your institution emphasize . . . providing the support you need to help you succeed academically? [envsuprt][b]
- To what extent does your institution emphasize . . . helping you cope with your non-academic responsibilities (work, family, etc.)? [envnacad][b]
- To what extent does your institution emphasize . . . providing the support you need to thrive socially? [envsocal][b]

Interpersonal Environment ($E\rho^2 = 0.80$)

- Quality of your relationships with . . . other students? [envstu][b]
- Quality of your relationships with . . . faculty members? [envfac]è:[b]
- Quality of your relationships with . . . administrative personnel and offices? [envadm][b]

continued

APPENDIX. *continued*

Outcome Measures

Gains in Practical Skills ($E\rho^2 = 0.74$)

- To what extent has you experience at this institution contributed to your knowledge, skills, and personal development in . . . using computing and information technology? [gncmpts]
- To what extent has you experience at this institution contributed to your knowledge, skills, and personal development in . . . analyzing quantitative problems? [gnquant]
- To what extent has you experience at this institution contributed to your knowledge, skills, and personal development in . . . acquiring job or work-related knowledge and skills? [gnwork]

Gains in General Education ($E\rho^2 = 0.81$)

- To what extent has you experience at this institution contributed to your knowledge, skills, and personal development in . . . writing clearly and effectively? [gnwrite]
- To what extent has you experience at this institution contributed to your knowledge, skills, and personal development in . . . speaking clearly and effectively? [gnspeak]
- To what extent has you experience at this institution contributed to your knowledge, skills, and personal development in . . . thinking critically and analytically? [gnanaly]
- To what extent has you experience at this institution contributed to your knowledge, skills, and personal development in . . . acquiring a broad general education? [gngenled]

[a] Generalizability coefficients are based on groups of 50 students.

[b] Items included in the NSSE benchmarks.

REFERENCES

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 19-32). Hillsdale, NJ: Lawrence Erlbaum.

Astin, A. W. (1984). Student involvement: A developmental theory for higher education. *Journal of College Student Personnel, 25,* 297-307.

Astin, A. W. (1985). Involvement: The cornerstone of excellence. *Change, 17*(4), 35-39.

Banta, T. W. (2002). A call for transformation. In T. Banta (Ed.), *Building a scholarship of assessment* (pp. 284-292). San Francisco: Jossey-Bass.

Banta, T. W., & Pike, G. R. (1989). Methods for evaluating assessment instruments. *Research in Higher Education, 30,* 455-470.

Brennan, R. L. (1983). *Elements of generalizability theory.* Iowa City, IA: ACT.

Campbell, D. T. (1960). Recommendations for APA test standards regarding construct, trait, or discriminant validity. *American Psychologist, 15,* 546-553.

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

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

El-Khawas, E. (2003, August). Using NSSE data for assessment and institutional improvement. *Documenting effective educational practices: National roundtable series* (Issue 5). Bloomington, IN: Indiana University Center for Postsecondary Research.

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. 75-126). Washington, DC: American Educational Research Association.

Ewell, P. T. (2002). An emerging scholarship: A brief history of assessment. In T. Banta (Ed.), *Building a scholarship of assessment* (pp. 3-25). San Francisco: Jossey-Bass.

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

Gellin, A. (2003). The effect of undergraduate student involvement on critical thinking: A meta-analysis of the literature, 1991-2000. *Journal of College Student Development, 44,* 746-762.

Indiana University Center for Postsecondary Research. (2004). *Student engagement: Pathways to collegiate success.* Bloomington, IN: Author.

Kane, M. T., Gillmore, G. M., & Crooks, T. J. (1976). Student evaluations of teaching: The generalizability of class means. *Journal of Educational Measurement, 13,* 171-183.

Kezar, A. (2002, December). Faculty developers using the National Survey of Student Engagement (NSSE) to be change agents. *Documenting effective educational practices: National roundtable series* (Issue 1). Bloomington, IN: Indiana University Center for Postsecondary Research.

Kezar, A. (2003, June). Student affairs administrators: Building collaborations with students and academic affairs for institutional improvement. *Documenting effective educational practices: National roundtable series* (Issue 3). Bloomington, IN: Indiana University Center for Postsecondary Research.

Kuh, G. D. (2001). *The National Survey of Student Engagement: Conceptual framework and overview of psychometric properties.* Bloomington, IN: Indiana University Center for Postsecondary Research.

Kuh, G. D. (2003). What we're learning about student engagement from NSSE. *Change, 35*(2), 24-32.

Kuh, G. D., Gonyea, R. M., & Palmer, M. (2001). The disengaged commuter student: Fact or fiction? *Commuter Perspectives, 27*(1), 2-5.

Kuh, G. D., Gonyea, R. M., & Rodriguez, D. P. (2002). The scholarly assessment of student development. In T. Banta (Ed.), *Building a scholarship of assessment* (pp. 100-128). San Francisco: Jossey-Bass.

Kuh, G. D., Hu, S., & Vesper, N. (2000). "They shall be known by what they do": An activities-based typology of college students. *Journal of College Student Development, 41,* 228-244.

Kuh, G. D., Schuh, J. H., Whitt, E. J., & Associates. (1991). *Involving colleges: Encouraging student learning and personal development through out-of-class experiences.* San Francisco: Jossey-Bass.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3,* 635-694.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Pace, C. R. (1980). Measuring the quality of student effort. *Current Issues in Higher Education, 2,* 10-16.

Pace, C. R. (1984). *Measuring the quality of college student experiences.* Los Angeles: Center for the Study of Evaluation, University of California Los Angeles.

Pascarella, E. T., Whitt, E. J., Nora, A., Edison, M., Hagedorn, L. S., & Terenzini, P. T. (1996). What have we learned from the first year of the national study of student learning? *Journal of College Student Development, 37,* 182-192.

Peterson, M. W., Einarson, M. K., Augustine, C. H., & Vaughan, D. S. (1999). *Institutional support for student assessment: Methodology and results of a national survey.* Stanford, CA: National Center for Postsecondary Improvement.

Pike, G. R. (1989). Background, college experiences, and the ACT-COMP exam: Using construct validity to evaluate assessment instruments. *Review of Higher Education, 13,* 91-118.

Pike, G. R. (1992). The components of construct validity: A comparison of two measures of general education outcomes. *Journal of General Education, 41,* 130-150.

Pike, G. R. (1994). Applications of generalizability theory in higher education assessment research. In J. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. X, pp. 45-87). New York: Agathon.

Pike, G. R. (1995). The relationship between self reports of college experiences and achievement test scores. *Research in Higher Education, 36,* 1-21.

Pike, G. R. (2002). Measurement issues in outcomes assessment. In T. Banta (Ed.), *Building a scholarship of assessment* (pp. 131-147). San Francisco: Jossey-Bass.

Pike, G. R., Kuh, G. D., & Gonyea, R. M. (2003). The relationship between institutional mission and students' involvement and educational outcomes. *Research in Higher Education, 44,* 243-263.

Pike, G. R. (2006). The dependability of NSSE Scalelets for college- and department-level assessment. *Research in Higher Education, 47,* 177-195.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Thissen, D., & Mislevy, R. L. (1990). Testing algorithms. In J. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 103-136). Hillsdale, NJ: Lawrence Erlbaum.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computer adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185-201.

Wainer, H., Dorans, N. J., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). Future challenges. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 233-271). Hillsdale, NJ: Lawrence Erlbaum.