Res High Educ (2013) 54:149–170
DOI 10.1007/s11162-012-9279-y

NSSE Benchmarks and Institutional Outcomes: A Note on the Importance of Considering the Intended Uses of a Measure in Validity Studies

Gary R. Pike

Received: 13 June 2012 / Published online: 7 December 2012

Abstract Surveys play a prominent role in assessment and institutional research, and the NSSE College Student Report is one of the most popular surveys of enrolled undergraduates. Recent studies have raised questions about the validity of the NSSE survey. Although these studies have themselves been criticized, documenting the validity of an instrument requires an affirmative finding regarding the adequacy and appropriateness of score interpretation and use. Using national data from NSSE 2008, the present study found that the NSSE benchmarks provided dependable means for 50 or more students and were significantly related to important institutional outcomes such as retention and graduation rates.

Keywords Surveys · Validity · Engagement · Retention · Graduation

Surveys of undergraduates' college experiences play a prominent role in assessment and institutional research. A recent survey of chief academic officers by the National Institute for Learning Outcomes Assessment (NILOA) found that 76 % of the respondents reported that their institutions made use of surveys in assessment (Kuh and Ikenberry 2009). The National Survey of Student Engagement's (NSSE) survey, *The College Student Report*, is among the most popular surveys of enrolled undergraduates, having been used by almost 1,500 institutions since 2000 (National Survey of Student Engagement 2011a). Several studies have documented the adequacy and appropriateness of using NSSE data for institution- and group-level decision making (Kuh 2001; Kuh et al. 2001, 2007; National Survey of Student Engagement 2010a, b, d, e, g, h, 2011b; Ouimet et al. 2004; Pascarella et al. 2009; Pike 2006a). In addition, a wide variety of institutions have identified

Paper presented at the annual meeting of the Association for Institutional Research, New Orleans, Louisiana, June 2012.

G. R. Pike (✉)
Indiana University-Purdue University-Indianapolis, 1100 Waterway Boulevard, WW53, Indianapolis, IN 46202, USA
e-mail: pikeg@iupui.edu

improvements to undergraduate education that have been made using NSSE data (Banta et al. 2009; Kuh 2005; National Survey of Student Engagement 2009, 2011b, c, f).

Recently, several studies have raised questions about the reliability and validity of NSSE and other surveys (Campbell and Cabrera 2011; DiRamio and Shannon 2010; Gordon et al. 2008; Korzekwa and Marley 2010; LaNasa et al. 2009; Lee 2010; Nora et al. 2011; Porter 2011; Porter et al. 2011). These studies have raised questions about the accuracy of students' self-reports, the structure of the NSSE benchmarks, and whether the benchmark scores are related to important educational outcomes. Although the studies critical of NSSE have themselves been the objects of criticism (see Ewell et al. 2011; McCormick and McClenney 2012), questions about the adequacy and appropriateness of self-report measures remain. Pike (2011) argued that a significant limitation of many of the studies focusing on the validity of survey data is that these studies have failed to consider the intended uses of the data.

Drawing on the validity framework of the American Educational Research Association, American Psychological Association, National Council for Measurement in Education (1999), as well as the work of Samuel Messick (1989) and Michael Kane (2006), the present research examined the adequacy and appropriateness of the NSSE benchmarks in light of their intended uses—institution-level decision making. This study advances knowledge about appropriate uses of the NSSE benchmarks in two important respects. First, the present research replicates the results of earlier studies (Pascarella et al. 2009; Pike 2006a) using data from a different point in time and using a much larger sample of institutions. Second, unlike the earlier studies, this research examines plausible rival hypotheses that could call into question conclusions about the adequacy and appropriateness of inferences and actions based on the NSSE benchmarks. As will be seen in the section that follows, measurement scholars have argued that replicating results and excluding rival hypotheses are essential elements in establishing the adequacy and appropriateness of measures such as the NSSE benchmarks.

Background

A Framework for Evaluating NSSE Benchmarks

The validity frameworks used to evaluate educational measures have evolved over time. Cureton (1951) proposed a standard based on criterion-related validity. In the case of the NSSE benchmarks, criterion validity would be represented by the relationship between benchmark scores and other measures of effective educational practice (e.g., student academic success). He argued that the appropriate framework for evaluating the validity of a measure is "the correlation between the actual test score and the 'true' test score" (Cureton 1951, p. 623). Criterion-related validity has been used extensively in educational and psychological research and is appropriate when a criterion variable is available; however, it is not an appropriate standard when the validity of the criterion measure is called into question (Kane 2006).

Recognizing the limitations of Cureton's approach, Cronbach (1971) proposed a variation on criterion-related validity in which a measure is compared to its specifications. In this case, the empirical structure of NSSE benchmarks would be compared to the technical specifications for the benchmarks. Cronbach's (1971) approach, which relied heavily on factor analysis to assess the structure of an instrument, is frequently used to evaluate achievement tests (Kane 2006), but is an inappropriate standard for evaluating educational measures that represent cognitive processes which are not readily observable (Cronbach 1971). It is also important to note that Cronbach (1971) argued that validity studies should propose and test plausible counter-hypotheses that call into question the validity of a measure. Even more significant, both Cureton's (1951) and Cronbach's (1971) validity frameworks presume that validity is a characteristic of the measure and do not allow researchers to evaluate how data are interpreted and used (Messick 1989).

Messick's (1989) construct-validity framework represented a significant departure from previous validity standards. He argued that validity judgments should focus on how data are interpreted and used. Messick (1989) defined validity as "an integrative and evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (p. 13). Based on this standard, a measure may be appropriate for one use, but not for another. Messick (1989) noted that replication is a critical element in validation, arguing that validity judgments should be based on the accumulation of evidence from a variety of studies, rather than on evidence from one study. Drawing on the work of Loevinger (1958), he argued that validity judgments should be based on an explicit theory of the constructs being measured and that validity studies should focus on the content and structure of the data, in addition to the relationships between those data and external measures. Thus, the content of the NSSE benchmarks should represent good educational practice, the structure of the benchmarks should be consistent with their specifications, and NSSE benchmarks should be related to measures of student success. Like Cronbach (1971), Messick (1989) stressed the importance of testing plausible rival hypotheses in validity studies. It is also noteworthy that Messick's (1989) definition of validity appears to have formed the basis for the validity framework set forth in the *Standards for Educational and Psychological Testing* by the American Educational Research Association, American Psychological Association, National Council for Measurement in Education (1999) p. 9):

> Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of the tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself.

Kane (2006) extended Messick's (1989) framework, emphasizing the need for an argument-based approach to validity. Kane (2006) also noted that it is inappropriate to construct a validity argument without first identifying how scores are to be interpreted and used: "The evidence needed for validation necessarily depends on the claims being made. Therefore, validation requires a clear statement of the proposed interpretations and uses" (p. 23). Drawing on the work of British logician Stephen Toulmin (1958), Kane (2006) argued that statements of intended uses and interpretations provide the warrants linking evidence and claim. He also argued that replication of validity studies is needed to identify the range of appropriate interpretations and uses of a measure, and that testing rival- or counter-hypotheses is essential in order to establish the limits of validity arguments. It is also significant to note that Kane (2006) argued that factor analysis can be an appropriate technique for generating validity evidence, but he recommended that factor analysis be used to evaluate relationships between a measure and external variables, rather than as a technique to evaluate the structural component of validity. He suggested that generalizability theory should be used to evaluate the structure of a measure.

The Interpretation and Use of NSSE Benchmarks

The concept of student engagement has its origins in the work of Ralph Tyler (1932), C. Robert Pace (1980, 1984), Alexander Astin (1984), Chickering and Gamson (1987), and Kuh et al. (1991). The rationale for assessing student engagement is based on two deceptively simple premises: (1) learning and success in college are related to the amount of time and effort students devote to educationally purposeful activities and (2) institutions can use their resources to influence the extent to which students are engaged in educational activities (Kuh 2003, 2006). NSSE's survey, *The College Student Report*, was developed to collect data from first-year students and seniors about their levels of engagement and their perceptions of institutional policies and priorities in order to document effective educational practices and identify opportunities for improving undergraduate education (Kuh 2001, 2009; Kuh et al. 2001).

In 2001, NSSE introduced five institutional benchmarks of effective educational practice: level of academic challenge, active and collaborative learning, student–faculty interaction, enriching educational experiences, and supportive campus environment (National Survey of Student Engagement 2001). The benchmarks were not intended to represent underlying theoretical constructs; instead, the benchmarks were conceived as clusters of student behaviors and institutional actions that represented good educational practices (Kuh 2001; Kuh et al. 2001). The purpose of the benchmarks was to concisely summarize an institution's educational practices and to facilitate conversations about improving undergraduate education (National Survey of Student Engagement 2001). The benchmarks also were intended to use institutions and/or groups of students as units of analysis (Kuh 2001; Kuh et al. 2001; National Survey of Student Engagement 2001). Noting that variation in student engagement is greater within than between institutions, NSSE researchers have cautioned decision makers against over-interpreting or acting on small differences in benchmark scores (Kuh 2007; National Survey of Student Engagement 2001).

Previous Research Supporting the NSSE Benchmarks

Research supporting the use of NSSE data for institutional assessment and improvement has focused on three issues paralleling Loevinger's (1957) content, structural, and external components of validity: (1) the clarity of NSSE questions and the consistency of student responses to NSSE items, (2) the dependability of institutional benchmark scores, and (3) the extent to which institutional/group benchmark scores are related to external (i.e., criterion) variables.

Content

NSSE staff members have relied heavily on focus groups and cognitive interviews to evaluate the clarity of NSSE survey questions and the consistency of students' responses to those questions. Early research revealed that students believed the questions were generally clearly worded and easy to understand. The students also had similar interpretations of the questions (Ouimet et al. 2004). Based on the early research, several questions were revised to improve interpretability. Subsequent studies confirmed earlier findings and also revealed that students from different racial/ethnic groups interpreted the NSSE survey questions in similar ways (Kuh et al. 2007; National Survey of Student Engagement 2010b, g). Students also reported they used different standards for selecting response options (e.g., never, sometimes, often, or very often) depending on the behavior described in the question. For example, very often generally meant daily for "asked questions in class," but meant three or four times a semester for "tutored other students" (Ouimet et al. 2004). Quantitative studies have confirmed these results and support the appropriateness of using vague quantifiers (e.g., never, seldom, often, and very often) to categorize students' behaviors, rather than asking students to provide precise counts of the frequency (e.g., once, twice, or ten times) of behaviors (National Survey of Student Engagement 2010d; Pace and Friedlander 1982; Wänke 2002; Wright et al. 1994). The results of these studies show that students are able to accurately and consistently report behaviors within broad categories, but are less accurate in making precise estimates of the amount of time spent on various activities.

Structure

Studies of the dependability of NSSE benchmark scores have focused on their appropriateness for institution- and group-level decision making. NSSE staff has reported that standard errors and sampling errors for the institutional benchmarks are within acceptable ranges (National Survey of Student Engagement 2010e). These researchers also reported that institutional benchmark scores were highly correlated from 1 year to the next (National Survey of Student Engagement 2011c). Pike (2006a) calculated group-mean generalizability coefficients using seniors' responses to the survey and found that the benchmarks produced highly reliable group means ($E\rho^2 \geq 0.70$) with as few as 50 students. Samples of 25 students produced group mean generalizability coefficients greater than 0.60.

External Relationships

Researchers have also examined the relationships between NSSE benchmarks and external measures. In early studies, Kuh et al. found significant differences in benchmark means based on students' academic majors (Kuh 2001; Kuh et al. 2001). More recently, NSSE staff compared benchmark means for known groups (e.g., first-year students vs. seniors, full-time vs. part-time students, and on-campus students vs. commuters) and found statistically significant differences in groups means that were in expected directions (National Survey of Student Engagement 2010a). NSSE researchers also found that benchmark scores were significantly and positively related to group-level persistence rates and credit hours earned (National Survey of Student Engagement 2010h).

Research has also documented that the NSSE benchmarks are appropriate for institution-level decision making. Pike (2006a), for example, reported that institutional benchmark scores were significantly and positively related to self-reported gains in general education and practical skills. Pascarella et al. (2009) found that benchmark scores were positively related to gains in institutional measures of critical thinking, moral development, openness to diversity, and psychological well-being. They concluded that "NSSE results regarding educational practices and student experiences are good proxy measures for growth in important educational outcomes" (Pascarella et al. 2009, p. 23).

Research Critical of the NSSE Benchmarks

Research critical of NSSE has also focused on similar issues related to content, structure, and external relationships: (1) the accuracy of students' self-reports in general (content), (2) the factor structure of the benchmarks (structure), and (3) the relationships between benchmark scores and educational outcomes at the student level (external relationships).

Content

Drawing on a rich literature from a variety of fields, Porter (2011) reviewed research concerning the accuracy of students' self-reports. He concluded that students have difficulty encoding and accurately reporting on their behaviors. Moreover, students tend to report in ways that put themselves in a positive (i.e., socially desirable) light. Porter et al. (2011) provided some empirical evidence concerning the inaccuracy of self-reports. They compared self-reports of the number of papers written to course syllabi for 42 students and found low rates (21 %) of exact agreement. More than half of the students over-reported the number of writing assignments in their classes.

Structure

Following Crohbach's (1971) example, at least four studies have attempted to replicate the structure of the benchmarks using either confirmatory (restricted) or exploratory (unrestricted) factor analysis. Campbell and Cabrera (2011) and LaNasa et al. (2009) conducted confirmatory factor analyses of NSSE items using data from single institutions and were unable to replicate the structure of the NSSE benchmarks. Both Lee (2010) and Nora et al. (2011) conducted factor analyses of data from the Community College Survey of Student Engagement (CCSSE) and were unable to replicate the structure of the CCSSE benchmarks which are similar to the NSSE benchmarks. Gordon et al. (2008) also raised questions about the NSSE benchmarks, noting that alpha reliability coefficients were frequently below 0.70. These reliabilities were not appreciably different from those reported by Kuh et al. (2007), although they were lower than recent reliability estimates based on national data (McCormick et al. 2009; Pike et al. 2011).

External Relationships

Several studies have examined the relationships between students' scores on the NSSE benchmarks and their educational outcomes. These single-institution studies failed to find statistically significant relationships between students' scores on the NSSE benchmarks and their grade point averages, persistence, or time to degree (Campbell and Cabrera 2011; DiRamio and Shannon 2010; Gordon et al. 2008; Korzekwa and Marley 2010). Similarly, NSSE researchers have reported weak, but statistically significant, relationships between NSSE engagement scores and standardized tests in national studies using students as the units of analysis (Carini et al. 2006).

Criticisms of Previous Studies

Many of the studies critical of the NSSE benchmarks have themselves been criticized for failing to take into account the intended uses of the benchmarks. For example, Porter's (2011) criticisms have been criticized for raising questions about whether students self-reports were precisely accurate, but failing to note that NSSE cautions against over interpreting small differences and encourages institutions to focus on major trends over time (Ewell et al. 2011; McCormick and McClenney 2012). Likewise, McCormick and McClenney (2012) criticized Porter (2011) for failing to address the fact that NSSE relies on vague quantifiers, rather than precise reports of behavior. McCormick and McClenney (2012) also criticized Porter (2011) for failing to respond to evidence from focus groups showing that NSSE respondents reported they understood the questions being asked and interpreted the questions in similar ways.

Studies of the factor structure of the NSSE benchmarks have been criticized for using inappropriate analytical methods. These criticisms are both conceptual and statistical. At the conceptual level, McCormick and McClenney (2012) argued that the items comprising the NSSE benchmarks are clusters of questions about related sets of good educational practices and were never intended to represent underlying psychological constructs. Because factor analysis assumes that the relationships among observed variables are the product of underlying constructs (Gorsuch 1983; McDonald 1985; Rummel 1970), it would not be appropriate for evaluating the structure of the NSSE benchmarks. The argument advanced by McCormick and McClenney (2012) is one of degree. To be sure, benchmark questions represent related sets of student behaviors. What is at question is whether the expected relationships would be strong enough to indicate the presence of underlying constructs, or whether the relationships are only strong enough to represent a dependable or generalizable indicator. There are also statistical issues with the use of factor analysis to evaluate the structure of the benchmark scores. Bernstein and Teng (1989) observed that factor analysis presumes that the data are continuous (i.e., scales), rather than a limited set of equal-appearing intervals (i.e., item responses). In their Monte Carlo study in which they knew the “true” underlying structure of the data, Bernstein and Teng (1989) found that factor analysis of Lykert-type items, such as those used by NSSE, failed to accurately represent the true structure of the data, producing instead spurious evidence of multidimensionality.

Studies of the relationships between students’ NSSE benchmark scores and their educational outcomes have been criticized for using inappropriate units of analysis. Because the benchmarks were designed to provide information about how colleges and universities foster student engagement, institutions and subgroups, not students, are the appropriate units of analysis (Ewell et al. 2011; McCormick and McClenney 2012). It is important to note that only institutional benchmark scores were available prior to 2004. Student-level scores were provided to allow institutions to calculate benchmarks for sub-groups (McCormick and McClenney 2012). According to NSSE supporters, studies using students as the units of analysis simply demonstrate that they survey should not be used in ways it was never intended to be used (Pike 2011).

Studies supporting the use of the NSSE benchmark scores for campus-level decision making can also be criticized. Studies by Pike (2006a) and by Pascarella et al. (2009) were limited to a single time period and relatively small samples of institutions. This is particularly true of the research by Pascarella et al. who limited their analyses to the 19 institutions participating in the Wabash National Study of Liberal Arts Education. In addition, these studies generally failed to examine other institutional characteristics/factors (i.e., rival hypotheses) that could have accounted for the positive relationships between institutions’ benchmark scores and their educational outcomes.

Research Questions

Simply criticizing research critical of the NSSE benchmarks does not provide support for the validity of the benchmarks. Documenting the validity of an educational measure requires affirmative findings about the adequacy and appropriateness of score interpretation and use. The present research examined the adequacy and appropriateness of using NSSE benchmark scores for institutional assessment and improvement. Two questions guided this research:

1. Do the NSSE benchmarks provide dependable measures for institutional and group-level decision making?
2. Are institutions' benchmark scores related to institution-level measures of student academic success?

The first research question focused on the structure of the NSSE benchmarks for first-year students and seniors. Given that the benchmarks were not intended to represent underlying constructs, factor analysis would not be appropriate. On the other hand, the benchmarks were developed to represent clusters of good educational practices that are conceptually related. To be useful, the benchmarks should provide dependable measures of good practice for groups of students. Generalizability theory was used in the present research to evaluate the structure of the NSSE benchmarks because it provides an appropriate framework for evaluating the dependability of group measures (Cronbach et al. 1972; Pike 1994). Generalizability theory represents an extension of classical reliability theory in that it relies on a multifaceted representation of measurement error (Feldt and Brennan 1989; Shavelson and Webb 1991). Generalizability theory assumes that an observed score represents responses to a sample of questions drawn from a universe of possible questions, and the generalizability coefficient provides an index of the dependability of generalizing from an observed score, based on a sample of questions or observations, to a mean score derived from the universe of acceptable observations (Cronbach et al. 1972).

Generalizability theory distinguishes between the *object of measurement*—that about which generalizations are to be made—and *facets of measurement*—the characteristics of the measurement situation that contribute to error (Brennan 1983). Generalizability theory partitions observed-score variance into variance that is attributable to the object of measurement and variance that is attributable to the facets of the measurement situation (Pike 1994). In generalizability theory, a distinction is made between *G* (generalizability) and *D* decision studies. *G* studies are designed to represent the universe of admissible observations and to provide estimates of the components of variance for that universe, whereas *D* studies represent the specifics of the measurement/research design and provide the basis for estimating the generalizability of an observed score (Brennan 1983). Kane (2006) identified two important advantages of generalizability theory when evaluating the structural component of validity. First, generalizability theory can identify the sources of error in a measurement situation so that strategies can be devised to minimize the effects of those aspects of the measurement situation. Second, generalizability analyses provide estimates of the standard errors of measurement and allow researchers to put limits on the precision of their estimates. Neither of these advantages accrues when factor analysis is used to assess the structural component of validity.

The second research question examined the relationships between institutional benchmarks and external measures of institutional quality and effectiveness, specifically IPEDS 1-year retention and 6-year graduation rates. This question addresses Cureton's (1951) concept of criterion-related validity. It goes beyond, traditional approaches to convergent-discriminant validity, however, by testing a series of rival hypotheses which posit that indicators of student success (i.e., retention and graduation rates are) are a product of institutional characteristics such as selectivity, mission, control, and the demographic characteristics of the student population. A finding that benchmark scores are related to institutional retention and graduation rates would suggest that the benchmarks can serve as proxies for institutional programs and practices that enhance student success above and beyond the characteristics of the institutions themselves.

Research Methods

Data Source

The data for this study came from the 2008 administration of the National Survey of Student Engagement. In 2008, almost one million first-year and senior students attending 722 4-year colleges and universities were invited to complete NSSE's survey, *The College Student Report.* Almost 380,000 students completed the survey—a response rate of 37 % (National Survey of Student Engagement 2008b). Several procedures were used to screen the institutions and students included in the analyses. First, students from special-mission (i.e., not research, Masters, or baccalaureate) institutions were excluded from the study. Next, students who took all of their courses via distance education were excluded from the study. Finally, institutions with fewer than 50 first-year or senior students were also excluded from the study. Using these procedures, 524 colleges and universities with 50 or more first-year respondents and 586 institutions with 50 or more senior respondents were selected for the study.

Table 1 presents the characteristics of the institutions included in the research. Not surprisingly, the characteristics of the institutions selected for the first-year and senior analyses were very similar. Approximately 55 % of the institutions were private, less than 20 % were research universities, almost 50 % were Master's universities, and approximately one-third were baccalaureate institutions. On average, nearly 60 % of the students attending the universities were female, less than 20 % were minority students, and nearly 90 % were full-time students. A comparison of benchmark scores for first-year and senior students revealed that the mean academic challenge, active and collaborative learning, student–faculty interaction, and enriching educational experiences benchmark scores were slightly higher for seniors, whereas the supportive campus environment benchmark scores were higher for first-year students. The average 1-year retention rate was nearly 78 %, and the average 6-year graduation rate was almost 55 %.

An examination of the fourth column in Table 1 reveals that the institutions included in this study are generally representative of all 4-year colleges and universities. The proportion of doctoral universities in the study is very similar to the proportion for all 4-year institutions, as are the measures of graduate coexistence, Barron's Selectivity Index, and the proportions of female and underrepresented minority students. The mean proportions of full-time students for the first-year and senior samples are slightly higher than for the population as a whole. Public Institutions and Master's institutions are slightly overrepresented in the sample and baccalaureate colleges are underrepresented. These differences, coupled with higher mean undergraduate enrollments for institutions included in the study, indicate a slight bias toward larger, public universities—institutions which educate the majority of students attending 4-year colleges and universities. An examination of the data

Table 1 Characteristics of the institutions included in the research

Institutional characteristic	First-year (N = 524)	Senior (N = 586)	4-Year institutions
Proportion of private institutions	0.55	0.55	0.67[d]
Proportion of doctoral institutions	0.19	0.17	0.18[d]
Proportion of Master's institutions	0.47	0.48	0.42[d]
Proportion of baccalaureate institutions	0.33	0.35	0.41[d]
Proportion of institutions with no graduate coexistence	0.23	0.25	0.28[b]
Proportion of institutions with some graduate coexistence	0.61	0.61	0.57[b]
Proportion of institutions with high graduate coexistence	0.15	0.14	0.15[b]
Barron's selectivity index	3.51	3.35	3.37[a]
Total undergraduate enrollment (in thousands)	5.84	5.52	3.72[c]
Proportion of female undergraduates	0.58	0.58	0.56[d]
Proportion of underrepresented minority undergraduates	0.16	0.18	0.23[d]
Proportion of full-time students	0.88	0.87	0.82[d]
Level of academic challenge	54.24	57.61	
Active and collaborative learning	44.15	52.81	
Student–faculty interaction	36.38	46.15	
Enriching educational experiences	28.67	43.83	
Supportive campus environment	63.10	60.35	
IPEDS retention rate (2007–2008)	77.13	NA	69.00[c]
IPEDS 6-year graduation rate (2007–2008, 2008–2009)	NA	54.86	48.50[c]

[a] Brint et al. (2011)

[b] Carnegie Foundation for the Advancement of Teaching (2010)

[c] National Center for Education Statistics (2012b)

[d] National Survey of Student Engagement (2008a)

for the outcomes measures used in the study also shows that the institutions included in the study had somewhat higher retention and graduation rates than the population of 4-year colleges and universities.

For the first part of the research, 50 institutions were randomly selected from the 524 institutions with first-year student scores. For each of these institutions, random samples of 50 students each were drawn from the total number of respondents. The same procedures were used to select seniors for this phase of the study. The end result was samples of 2,500 first-year students and 2,500 seniors. There were both conceptual and statistical reasons for the use of random sampling to create a balanced design for the *G* study. Conceptually, generalizability theory assumes that the objects of measurement, institutions in this study, are randomly drawn from the population of institutions (Pike 2006b). Likewise, the random selection of students within institutions was used because it is a widely accepted method of generating a representative sample of what served as a facet of measurement in the *G* study (students). Samples of 50 students each from 50 institutions were selected because a balanced design greatly simplifies the calculation of variance components in a *G* study. Owing to the random selection process, there was almost no overlap in the institutions selected for the two sets of generalizability analyses. For the second phase of the study, data for the first-year and senior students were aggregated at the institution level and merged with measures of institutional characteristics.

Measures

The measures used in the generalizability analyses were the 42 items comprising the NSSE benchmarks: level of academic challenge (11 items), active and collaborative learning (seven items), student–faculty interaction (six items), enriching educational experiences (12 items), and supportive campus environment (six items). A list of the items and their corresponding benchmarks is included in the Appendix. Because the number of response options differed for some items, all item scores were placed on a 0–100 scale using procedures recommended by the survey developer (National Survey of Student Engagement 2011b).

The measures used in the second phase of the data analysis were aggregated at the institution level. These measures included institutional benchmark scores for first-year students and seniors, IPEDS 1-year retention rates for 2007–2008, and mean IPEDS 6-year graduation rates based on data from 2007 to 2008 and 2008 to 2009. An average 6-year graduation rate was calculated because some of the seniors surveyed in 2008 would not be expected to graduate until 2008–2009. In addition to the NSSE and IPEDS measures, several institutional characteristics were included as controls. These variables have been found to be significantly related to student engagement and institutional outcomes (Astin and Oseguera 2005; Gansemer-Topf and Schuh 2006; Melguizo 2008; McCormick et al. 2009; Pike et al. 2006; Ryan 2004). Most important, these institutional characteristics served as plausible rival hypotheses in the validity study. Significant relationships between institutional characteristics and retention and graduation, as opposed to significant relationships between the NSSE benchmarks and retention and graduation, would indicate that any observed correlations between benchmark scores and student-success outcomes are a spurious result of relationships between institutional characteristics and the NSSE benchmarks.

Institutional (public vs. private) control was dichotomously scored to represent private institutions, whereas institutional mission was represented by two dummy variables—doctoral and Master's universities. Baccalaureate institutions served as the reference group for the institutional-mission measures. Graduate coexistence, the percentage of undergraduate degree programs with a corresponding graduate degree program, was also represented by two dummy variables—some graduate coexistence and high graduate coexistence. Institutions with little or no graduate coexistence served as the reference group. Several measures were derived from IPEDS institutional characteristics. Institutional size was represented by total undergraduate enrollment in thousands. Other measures included the proportion of female undergraduates, the proportion of underrepresented minority undergraduates, and the proportion of full-time undergraduates attending the institution. Barron's selectivity index was also included as an institutional characteristic.

Data Analysis

Separate generalizability analyses were conducted for first-year students and seniors using the BMDP8V computer program (Dixon 1992). First, variance components representing the elements (i.e., facets) of the measurement situation were calculated through a *G* study. The variance components were then used in a *D* study to estimate group mean generalizability coefficients (Shavelson and Webb 1991). The facets of measurement in the study were universities (*U*), items (*I*), students within universities ($S|U$), the university–item interaction (*UI*), and the item–student within university interaction ($IS|U$). In the *D* study, generalizability coefficients for group/institutional means were calculated using procedures recommended by Kane et al. (1976), and the formula for generalizing over students, not items, was utilized because the items comprising the NSSE benchmarks do not represent samples from an underlying construct or domain. Generalizability coefficients were calculated for samples of 25, 50, and 100 students.

The second phase of the data analysis examined the relationships between NSSE benchmarks and institutional retention and graduation rates, net the effects of other institutional characteristics. Benchmark scores for first-year students were compared to IPEDS retention rates, and benchmark scores for seniors were compared to average 6-year graduation rates. Stata 10 correlation and multiple regression procedures were used in the analyses (StataCorp 2007). Preliminary analyses indicated that the homogeneity of variance assumption was not met (Cook and Weisberg 1983). As a consequence, robust standard errors appropriate for heteroskedasticity were utilized.

Results

Generalizability Analyses

Table 2 presents the *D* study group mean generalizability coefficients for first-year and senior students. Not shown, but available from the author, are the *G* study variance components used to calculate the generalizability coefficients. An examination of the generalizability coefficients for first-year students reveals that only the Level of Academic Challenge and Active and Collaborative Learning benchmarks produced acceptable levels of dependability ($E\rho^2 \geq 0.70$) when the means were based on 25 students. When benchmark scores were based on 50 first-year students, all of the benchmarks, except Supportive Campus Environment, produced satisfactory generalizability coefficients. Even the Supportive Campus Environment benchmark's generalizability coefficient closely approached acceptable levels of dependability. When the first-year benchmark scores were based on 100 students, all of the group mean generalizability coefficients exceeded 0.80.

Table 2 D-study generalizability coefficients for first-year and senior students

First-year students

NSSE benchmark	Eρ^2 (N = 25)	Eρ^2 (N = 50)	Eρ^2 (N = 100)
Level of academic challenge	0.74	0.85	0.92
Active and collaborative learning	0.71	0.83	0.91
Student–faculty interaction	0.54	0.70	0.82
Enriching educational experiences	0.63	0.78	0.87
Supportive campus environment	0.51	0.68	0.81

Senior students

NSSE benchmark	Eρ^2 (N = 25)	Eρ^2 (N = 50)	Eρ^2 (N = 100)
Level of academic challenge	0.55	0.71	0.83
Active and collaborative learning	0.60	0.75	0.86
Student–faculty interaction	0.57	0.72	0.84
Enriching educational experiences	0.82	0.90	0.95
Supportive campus environment	0.63	0.77	0.87

Generalizability results for the senior benchmarks also revealed that group means based on 25 students were not very dependable. Only the generalizability coefficient for the enriching educational experiences benchmark exceeded the 0.70 threshold. In contrast, all five of the senior benchmarks produced acceptable levels of dependability ($E\rho^2 \geq 0.70$) when based on as few as 50 students. Group mean generalizability coefficients for the senior benchmark scores were all greater than 0.80 when group means were based on 100 students.

NSSE Benchmarks and Academic Success

Table 3 presents the results of the correlation and regression analyses for first-year students and seniors. An examination of the results for first-year students reveals that all of the institutional characteristics and four of the five benchmarks were significantly correlated with institutions' 1-year retention rates. However, only six measures were significantly related to institutional retention rates, net the effects of the other variables in the model. Barron's Selectivity Index, total undergraduate enrollment, and the proportion of undergraduate students who were enrolled full-time at an institution were positively related to retention rates. Two of the NSSE benchmark scores, level of academic challenge and supportive campus environment, were positively related to institutional retention rates, whereas the student–faculty interaction benchmark was negatively related to retention rates. It is also important to note that the level of academic challenge benchmark had the second strongest relationship with institutional retention rates, net the effects of the other variables in the model. The multiple regression analysis produced a R^2 coefficient of 0.69, indicating that 69 % of the variance in institutions' 1-year retention rates could be accounted for by the variables in the model.

Table 3 Correlation and regression results for first-year and senior students

	First-year students		Senior students	
	r_{xy}	β	r_{xy}	β
Private institution	0.14*	−0.06	0.35*	0.12*
Doctoral institution	0.21*	0.08	0.15*	0.05
Master's institution	−0.26*	0.09	−0.28*	0.04
Some graduate coexistence	−0.27*	0.01	−0.21*	0.04
High graduate coexistence	0.21*	−0.02	0.12*	0.02
Barron's selectivity index	0.75*	0.41*	0.76*	0.40*
Total enrollment (in thousands)	0.17*	0.17*	0.02	0.13*
Proportion female	−0.16*	−0.04	−0.12*	−0.01
Proportion underrepresented minority	−0.26*	−0.01	−0.44*	−0.16*
Proportion full-time	0.45*	0.16*	0.61*	0.25*
Level of academic challenge	0.52*	0.39*	0.51*	0.15*
Active and collaborative learning	−0.02	−0.05	0.01	−0.07
Student–faculty interaction	−0.15*	−0.26*	0.31*	−0.16*
Enriching educational experiences	0.38*	0.04	0.62*	0.22*
Supportive campus environment	0.24*	0.14*	0.23*	0.11*
Squared multiple correlation (R^2)		0.69		0.75

Results for seniors revealed that all of the institutional characteristics, except total undergraduate enrollment, were significantly correlated with mean 6-year graduation rates. Once again, four of the five NSSE benchmarks were positively correlated with the graduation-rate measure. An examination of the standardized regression coefficients in Table 3 shows that five institutional characteristics were significantly related to institutional graduation rates. Being a private institution, Barron's Selectivity Index, total undergraduate enrollment, and the proportion of full-time undergraduates at the institution were positively related to mean graduation rates, although the effect for total undergraduate enrollment should be viewed with skepticism because the correlation between total enrollments and graduation rates was not statistically significant. The proportion of underrepresented minority undergraduates at an institution was negatively related to the institution's average graduation rate.

The R^2 coefficient for the regression model was 0.75, indicating that 75 % of the variance in institution's average 6-year graduation rates was accounted for by the variables in the model. Examination of the standardized regression coefficients for the senior benchmark scores reveals that the level of academic challenge, enriching educational experiences, and supportive campus environment benchmarks were significantly and positively related to institutions' mean graduation rates, net the effect of the other variables in the model. Seniors' scores for the student–faculty interaction benchmark were negatively related to institutional graduation rates; however, this relationship should be interpreted with caution given that the correlation between student–faculty interaction scores and mean graduation rates was positive and statistically significant. It is also important to note that the enriching educational experiences benchmark had the third strongest association with graduation rates.

Limitations

Care should be taken not to over generalize the results of this research. These results are based on data from institutions that participated in the 2008 administration of the NSSE survey. Although NSSE results for a given year are very similar to results for other years, using data from a different survey administration could have produced different results. Also, the results reported in this study are based on a subset of the institutions participating in NSSE 2008. Because special-mission institutions and institutions with less than 50 respondents were not included in the study, care should be taken when extending the findings to those institutions. Because the focus of this study was on the NSSE benchmarks, the results should not be generalized to other surveys. Likewise, the present research examined the use of the NSSE benchmarks for assessment, and it is not appropriate to use these findings to justify other uses of the survey.

Perhaps the greatest limitation of this study is the criterion measure used to evaluate the NSSE benchmarks—IPEDS retention and graduation rates. Adelman and others have been critical of IPEDS retention and graduation rates because they focus on what is in some instances a very limited subsample of students (first-time, full-time, degree-seeking beginners). The IPEDS measures also fail to account for students who leave an institution, but remain in higher education and finish elsewhere (Adelman 1999, 2007; National Center for Public Policy and Higher Education 2002). To be sure, IPEDS retention and graduation rates present a limited view of institutional quality and effectiveness; nevertheless, the IPEDS measures are general indicators of student success at an institution (Hagedorn 2005). These retention and graduation rates also form the basis for a variety of policy indicators, such as "Student Right to Know" (National Center for Education Statistics

2012a), and the measures are widely used in research on institutional effectiveness (Gansemer-Topf and Schuh 2006; Ryan 2004).

Discussion

The findings of the current research can be summarized as follows:

1. The results demonstrated that the NSSE benchmarks can produce dependable measures of student engagement in good educational practices with as few as 50 students. With one exception, the group mean generalizability coefficients for first-year and senior students exceeded accepted standards for the dependability of educational measures ($E\rho^2 \geq 0.70$). When group means were based on 100 students, dependability coefficients were all greater than 0.80.
2. Multiple regression results clearly indicated that the NSSE institutional benchmark scores are significantly related to institutional retention and graduation rates, net the effects of institutional characteristics. In fact, NSSE benchmark scores were among the factors that were most strongly related to retention and graduation rates.

Despite their limitations, the findings of the present research have important implications for theory and practice in assessment and institutional research. First and foremost, the results indicate that the NSSE benchmarks can be used to assess the extent to which an institution's first-year and senior students are engaged in educationally purposeful activities. Because institutional policies and practices influence student engagement, the NSSE benchmarks can provide measures of the extent to which colleges and universities are effective in facilitating student engagement. Furthermore, this research suggests that assessment and institutional research professionals are not limited to using NSSE results only for institutional assessment. The fact that the benchmarks produce dependable means with as few as 50 students indicates that it is possible to use benchmark scores to gauge the engagement of student subgroups and evaluate the effectiveness of institutional actions and focused programs to improve student engagement and academic success. The caveat is that analyses should be based on groups of 50 or more students and that care should be taken not to over interpret small score differences.

This research also has implications for evaluating the validity of assessment measures and for selecting measures based on extant validity studies. One lesson to be learned from the current study is the importance of clearly defining the intended uses of assessment measures. Although this recommendation seems obvious, it is not always followed. Collecting and reporting data about unintended uses of a measure (e.g., student-level diagnosis in the case of NSSE) is a waste of effort and can confuse decisions about the adequacy and appropriateness of an assessment instrument. Conversely, collecting and evaluating data about the appropriateness of an instrument for a clearly defined use, such as institution or program assessment and evaluation, can help define the conditions under which assessment measures are likely to yield accurate and appropriate information for improvement.

Defining the intended uses of a measure, as well as the assumptions underlying the measure, also influences the methods used to evaluate an instrument. Several studies have used factor analysis to evaluate the structure of the NSSE benchmarks. This approach was based on the assumption that the benchmarks are scales representing underlying psychological constructs. However, the survey's developers designed the benchmarks to represent clusters of good educational practices and to provide a starting point for examining specific aspects of student engagement (Ewell et al. 2011; Kuh 2001; McCormick and McClenney

2012). Given the nature and intended uses of the benchmarks, generalizability analyses, rather than factor analysis, is most appropriate. In addition, because the items comprising the benchmarks are not assumed to be random samples from universes of acceptable items, an analysis of generalizability over students, but not items, is most appropriate.

Closely related to the need to define the intended uses of an instrument is the need to select appropriate units of analysis. The NSSE benchmarks are a case in point. Studies critical of the NSSE benchmarks used students as the units of analysis and found little relationship between students' scores and retention and graduation. A very different picture emerged when institutions were the unit of analysis. Institutional benchmarks were strongly related to IPEDS retention and graduation rates. As Ewell (1991) noted, group-level measures are appropriate for evaluation/self-study and accountability/quality assurance, whereas student-level measures are appropriate for diagnosis and certification, Thus, the NSSE benchmarks are appropriate for assessment and evaluation, but not for evaluating or predicting the academic success of individual students.

The results of this research also underscore the importance of considering rival hypotheses in validity studies. As Cronbach (1971), Messick (1989), and Kane (2006) observed, including rival hypotheses in validity studies helps ensure that findings supporting the adequacy and appropriateness of an interpretation or use of data are not the spurious result of the relationships between the measure being evaluated and some other (external) measure. In the present research, multiple regression results clearly demonstrate that the good educational practices represented by the NSSE benchmarks are related to the success of institutions in promoting student academic success (i.e., retention and graduation rates), above and beyond the characteristics of the institutions.

Including rival hypotheses can also identify limitations that should be placed on proposed interpretations and uses (Cronbach 1971; Kane 2006; Messick 1989). The findings of this study clearly show a close linkage between institutional selectivity and retention and graduation rates. Barron's Selectivity Index was the institutional characteristic most strongly related to both IPEDS retention and graduation rates, net the effects of other variables including the NSSE benchmarks. This finding is consistent with the results of previous research on the relationships between institutional characteristics and student success (Astin and Oseguera 2005; Gansemer-Topf and Schuh 2006; Melguizo 2008). In addition, institutional size (total undergraduate enrollment) and the proportion of full-time students were positively related to institutional retention and graduation rates, whereas the proportion of underrepresented minority students at an institution was negatively related to graduation rates. Again, many of these findings are consistent with previous research (Astin and Oseguera 2005; Ryan 2004). In the context of this validity study, these results suggest that student engagement, as represented by the NSSE benchmarks, can and does enhance student success across a wide range of different 4-year institutions, but, on average, retention and graduation rates will still be higher for more selective institutions, larger institutions, and institutions with higher proportions of full-time students.

The findings of the present research also have implications for theory and research related to student engagement. Results revealed that both the academic challenge and campus environment measures were significantly and positively related to 1-year retention and 6-year graduation rates. It would appear that student success in college is facilitated by institutions setting high academic standards and then supporting students as they work to meet those standards. This finding is consistent with Chickering and Gamson's (1987) "Principles of Good Practice in Undergraduate Education." It is also consistent with the recommendations in *Involvement in Learning* by the National Institute of Education's Study Group on the Conditions of Excellence in American Higher Education (1984).

The negative relationships between student–faculty interaction scores and IPEDS retention and graduation rates are surprising, although the presence of a significant positive correlation between institutional student–faculty interaction benchmark scores for seniors and 6-year graduation rates suggests that the significant negative beta in the regression analysis is a statistical artifact. However, both the correlation and regression coefficients representing the relationships between the student–faculty interaction benchmarks for first-year students and IPEDS 1-year retention rates were negative and statistically significant. It may be tempting to think that high levels of faculty-student interaction lead to lower retention rates, but another interpretation is possible. Perhaps faculty members at institutions where retention is a problem spend more time interacting with students in an effort to help those students be successful academically.

It is also interesting to note that scores on the enriching educational experiences benchmark were significantly related to institutions' 6-year graduation rates. In fact, enriching educational experiences scores were the third strongest factor explaining institutional graduation rates. This finding appears to contradict the conventional wisdom of many outside higher education that students' progress toward a degree can be slowed when they become involved in a variety of activities outside the classroom. Instead, involvement in a variety of educational experiences including internships, self-designed majors, and study abroad, appear to increase the likelihood of receiving a degree in 6 years.

Conclusion

Peter Ewell (1991) observed that assessment for improvement and accountability requires data about students' experiences and outcomes aggregated at the institutional or group levels, not student-level data. Although some previous studies have raised questions about the utility of the NSSE benchmarks for predicting the academic success of individual students, the present research found that the benchmarks provide dependable measures that are related to important indicators of quality and effectiveness at the institution level. The results of this study also underscore the importance of clearly identifying how data are to be interpreted and used before undertaking a validity study. As Messick (1989), Kane (2006), and the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, National Council for Measurement in Education 1999) observed, a measure can be valid for one interpretation or use, but not for another. Based on the results of this study, the NSSE benchmarks appear to be adequate and appropriate measures of student engagement for the purposes of assessment and evaluation but not for the purposes of diagnosis or certification.

Appendix

Items Comprising the NSSE Benchmarks

Level of Academic Challenge

- Preparing for class (studying, reading, writing, rehearsing etc. related to academic program)
- Number of assigned textbooks, books, or book-length packs of course readings
- Number of written papers or reports of 20 pages or more

- Number of written papers or reports of between 5 and 19 pages
- Number of written papers or reports of fewer than five pages
- Coursework emphasizing analysis of the basic elements of an idea experience or theory
- Coursework emphasizing synthesis and organizing of ideas, information, or experiences into new, more complex interpretations and relationships
- Coursework emphasizing the making of judgments about the value of information, arguments, or methods
- Coursework emphasizing application of theories or concepts to practical problems or in new situations
- Working harder than you thought you could to meet an instructor's standards or expectations
- Campus environment emphasizing time studying and on academic work

Active and Collaborative Learning

- Asked questions in class or contributed to class discussions
- Made a class presentation
- Worked with other students on projects during class
- Worked with classmates outside of class to prepare class assignments
- Tutored or taught other students
- Participated in a community-based project as part of a regular course
- Discussed ideas from your readings or classes with others outside of class (students, family members, co-workers, etc.)

Student–Faculty Interaction

- Discussed grades or assignments with an instructor
- Talked about career plans with a faculty member or advisor
- Discussed ideas from your readings or classes with faculty members outside of class
- Worked with faculty members on activities other than coursework (committees, orientation, student-life activities, etc.)
- Received prompt feedback from faculty on your academic performance (written or oral)
- Worked with a faculty member on a research project outside of course or program requirements

Enriching Educational Experiences

- Participating in co-curricular activities (organizations, publications, student government, sports, etc.)
- Practicum, internship, field experience, co-op experience, or clinical assignment
- Community service or volunteer work
- Foreign language coursework
- Study abroad
- Independent study or self-designed major
- Culminating senior experience (comprehensive exam, capstone course, thesis, project, etc.)

- Serious conversations with students of different religious beliefs, political opinions, or personal values
- Serious conversations with students of a different race or ethnicity
- Using electronic technology to discuss or complete an assignment
- Campus environment encouraging contact among students from different economic, social, and racial or ethnic backgrounds
- Participate in a learning community or some other formal program where groups of students take two or more classes together

Supportive Campus Environment

- Campus environment provides the support you need to help you succeed academically
- Campus environment helps you cope with your non-academic responsibilities (work, family, etc.)
- Campus environment provides the support you need to thrive socially
- Quality of relationships with other students
- Quality of relationships with faculty members
- Quality of relationships with administrative personnel and offices

References

Adelman, C. (1999). *Answers in the tool box: Academic intensity, attendance patterns, and bachelor's degree attainment, report PLLI* (pp. 1999–8021). Washington, D.C.: U. S. Department of Education.

Adelman, C. (2007). Making graduation rates matter. In *Inside higher education.* Retrieved February 3, 2012 from http://www.insidehighered.com/views/2007/03/12/adelman.

American Educational Research Association, American Psychological Association, National Council for Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, D.C.: American Educational Research Association.

Astin, A. W. (1984). Student involvement: A developmental theory for higher education. *Journal of College Student Development, 25*, 297–308.

Astin, A. W., & Oseguera, L. (2005) Pre-college and institutional influences on degree attainment In A. Seidman (Ed.), *College student retention: Formula for student success* (pp. 245–276). Westport, CT: American Council on Education and Praeger.

Banta, T. W., Pike, G. R., & Hansen, M. J. (2009). The use of engagement data in accreditation, planning and assessment. In R. M. Gonyea & G. D. Kuh (Eds.), *Using NSSE in institutional research* (new directions for institutional research series, no. 141, pp. 21–34). San Francisco: Jossey-Bass.

Bernstein, I. H., & Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin, 105*, 467–477.

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

Brint, S., Mulligan, K., Rotondi, M. B., & Apkarian, J. (2011). *The institutional data archive on American higher education, 1970–2010.* Riverside, CA: University of California. Retrieved from http://www.higher-ed2000.ucr.edu/databases.html/.

Campbell, C. M., & Cabrera, A. F. (2011). How sound is NSSE? Investigating the psychometric properties of NSSE at a public, research-extensive institution. *Review of Higher Education, 35*, 77–103.

Carini, R. M., Kuh, G. D., & Klein, S. P. (2006). Student engagement and student learning: Testing the linkages. *Research in Higher Education, 47*, 1–32.

Carnegie Foundation for the Advancement of Teaching (2010). *Summary tables: Undergraduate instructional program classification.* Stanford, CA: Author. Retrieved September 12, 2010 from http://classifications.carnegiefoundation.org/summary/ugrad_prog.php.

Chickering, A. W., & Gamson, Z. F. (1987). Seven principles for good practice in undergraduate education. *AAHE Bulletin, 39*(7), 3–7.

Cook, R. D., & Weisberg, S. (1983). Diagnostics for heteroskedasticity in regression. *Biometrika, 70*, 1–10.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, D.C.: 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.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–694). Washington, D.C.: American Council on Education.

DiRamio, D., & Shannon, D. (2010, April). *Is NSSE messy? An analysis of predictive validity*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Dixon, W. J. (1992). *BMDP statistical software manual*. Berkeley: University of California Press.

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, D.C.: American Educational Research Association.

Ewell, P. T., McClenney, K., & McCormick, A. C. (2011). Measuring engagement. In *Inside higher education*. Retrieved September 20, 2011 from http://www.insidehighered.com/views/2011/09/20/essay_defending_the_value_of_surveys_of_student_engagement.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: American Council on Education and Macmillan.

Gansemer-Topf, A. M., & Schuh, J. H. (2006). Institutional selectivity and institutional expenditures: Examining organizational factors that contribute to retention and graduation. *Research in Higher Education, 47*, 614–641.

Gordon, J., Ludlum, J., & Hoey, J. J. (2008). Validating NSSE against student outcomes: Are they related? *Research in Higher Education, 49*, 19–39.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale: Lawrence Erlbaum Associates.

Hagedorn, L. S. (2005). How to define retention: A new look at an old problem. In A. Seidman (Ed.), *College student retention: A formula for student success* (pp. 89–106). Westport: American Council on Education and Praeger.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport: American Council on Education and Praeger.

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.

Korzekwa, A. M., & Marley, S. C. (2010, April). *An examination of the predictive validity of National Survey of Student Engagement benchmarks and scalelets*. Paper Presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Kuh, G. D. (2001). *The national survey of student engagement: Conceptual framework and overview of psychometric properties*. Bloomington: 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. (2005). Putting student engagement results to use: Lessons from the field. *Assessment Update: Progress, Trends, and Practices in Higher Education, 17*(1), 12–13.

Kuh, G. D. (2006). Making students matter. In J. C. Burke (Ed.), *Fixing the fragmented university: Decentralization with discretion* (pp. 235–264). Boston: Jossey-Bass.

Kuh, G. D. (2007). Risky business: Promise and pitfalls of institutional transparency. *Change, 39*(5), 30–35.

Kuh, G. D. (2009). The national survey of student engagement: Conceptual and empirical foundations. In R. M. Gonyea & G. D. Kuh (Eds.), *Using NSSE in institutional research* (new directions for institutional research series, no. 141, pp. 5–20). San Francisco: Jossey-Bass.

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.

Kuh, G. D., Hayek, J. C., Carini, R. M., Ouimet, J. A., Gonyea, R. M., & Kennedy, J. (2001). *NSSE technical and norms report*. Bloomington: Indiana University Center for Postsecondary Research.

Kuh, G. D., & Ikenberry, S. (2009). *More than you think, less than we need: Learning outcomes assessment in American higher education*. Champaign: National Institute for Learning Outcomes Assessment.

Kuh, G. D., Kinzie, J., Cruce, T., Shoup, R., & Gonyea, R. M. (2007). *Connecting the dots: Multifaceted analyses of the relationships between student engagement results from the NSSE, and the institutional practices and conditions that foster student success*. Bloomington: Final report prepared for Lumina Foundation for Education. Center for Postsecondary Research.

LaNasa, S. M., Cabrera, A. F., & Transgrud, H. (2009). The construct validity of student engagement: A confirmatory factor analysis approach. *Research in Higher Education, 50*, 315–332.

Lee, C. (2010, April). *The reliability of national benchmarks of effective student engagement*. Paper Presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3,* 635-694 (Monograph Supplement 9).

McCormick, A. C., & McClenney, K. (2012). Will these trees *ever* bear fruit? A response to the special issue on student engagement. *Review of Higher Education, 35*, 307–333.

McCormick, A. C., Pike, G. R., Kuh, G. D., & Chen, D. P. (2009). Comparing the utility of the 2000 and 2005 Carnegie classification systems in research on students' college experiences and outcomes. *Research in Higher Education, 50*, 144–167.

McDonald, R. P. (1985). *Factor analysis and related methods*. Hillsdale: Lawrence Erlbaum Associates.

Melguizo, T. (2008). Quality matters: Assessing the impact of attending more selective institutions on college completion rates of minorities. *Research in Higher Education, 49*, 214–236.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan.

National Center for Education Statistics (2012a). *About IPEDS.* Retrieved March 20, **2012** from http://nces.ed.gov/ipeds/about/.

National Center for Education Statistics (2012b). *IPEDS data center.* Retrieved from http://nces.ed.gov/ipeds/datacenter/.

National Center for Public Policy and Higher Education (2002). *The different dimensions of transfer.* Retrieved February 3, 2012 from http://www.highereducation.org/reports/transfer/transfer5.shtml.

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.

National Survey of Student Engagement. (2001). *Improving the college experience: National benchmarks of effective educational practice*. Bloomington: Indiana University Center for Postsecondary Research.

National Survey of Student Engagement. (2008a). *NSSE 2008 overview.* Bloomington, IN: Indiana University Center for Postsecondary Research. Retrieved from http://nsse.iub.edu/pdf/2008_Institutional_Report/NSSE2008Overview.pdf.

National Survey of Student Engagement. (2008b). *Promoting engagement for all students: The imperative to look within*. Bloomington: Indiana University, Center for Postsecondary Research.

National Survey of Student Engagement. (2009). *Using NSSE to assess and improve undergraduate education: Lessons from the field, 2009*. Bloomington: Indiana University Center for Postsecondary Research.

National Survey of Student Engagement (2010a). *2009 Known groups validation.* Bloomington, IN: Indiana University Center for Postsecondary Research. Retrieved August 14, 2011 from http://www.nsse.iub.edu/links/psychometric_portfolio.

National Survey of Student Engagement (2010b). *Cognitive interviews and focus groups.* Bloomington, IN: Indiana University Center for Postsecondary Research. Retrieved August 14, 2011 from http://www.nsse.iub.edu/links/psychometric_portfolio.

National Survey of Student Engagement (2010c). *Consequential aspect of validity.* Bloomington, IN: Indiana University Center for Postsecondary Research. Retrieved August 14, 2011 from http://www.nsse.iub.edu/links/psychometric_portfolio.

National Survey of Student Engagement (2010d). *Do different versions of NSSE questions produce the "same" or similar results? Specifically, how often is often?* Bloomington, IN: Indiana University Center for Postsecondary Research. Retrieved August 14, 2011 from http://www.nsse.iub.edu/links/psychometric_portfolio.

National Survey of Student Engagement (2010e). *Do institutions participating in NSSE have enough respondents to adequately represent their population?* Bloomington, IN: Indiana University Center for Postsecondary Research. Retrieved August 14, 2011 from http://www.nsse.iub.edu/links/psychometric_portfolio.

National Survey of Student Engagement (2010f). *Do institutions use survey data as intended by NSSE?* Bloomington, IN: Indiana University Center for Postsecondary Research. Retrieved August 14, 2011 from http://www.nsse.iub.edu/links/psychometric_portfolio.

National Survey of Student Engagement (2010g). *Focus groups.* Bloomington, IN: Indiana University Center for Postsecondary Research. Retrieved August 14, 2011 from http://www.nsse.iub.edu/links/psychometric_portfolio.

National Survey of Student Engagement (2010h). *Predicting retention and degree progress.* Bloomington, IN: Indiana University Center for Postsecondary Research. Retrieved August 14, 2011 from http://www.nsse.iub.edu/links/psychometric_portfolio.

National Survey of Student Engagement (2011a). *About NSSE.* Bloomington, IN: Indiana University Center for Postsecondary Research. Retrieved August 13, 2011 from http://www.nsse.iub.edu/html/about.cfn.

National Survey of Student Engagement (2011b). *Does the NSSE survey produce similar results when administered to different cohorts of students at the same institutions across consecutive years?*

Bloomington, IN: Indiana University Center for Postsecondary Research. Retrieved August 14, 2011 from http://www.nsse.iub.edu/links/psychometric_portfolio.

Nora, A., Crisp, G., & Matthews, C. (2011). A reconceptualization of CCSSE's benchmarks of student engagement. *Review of Higher Education, 35*, 105–130.

Ouimet, J. A., Bunnage, J. B., Carini, R. M., Kuh, G. D., & Kennedy, J. (2004). Using focus groups to establish the validity and reliability of a college student survey. *Research in Higher Education, 45*, 233–250.

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. An account of the development and use of the college student experiences questionnaire*. Los Angeles: Higher Education Research Institute.

Pace, C. R., & Friedlander, J. (1982). The meaning of response categories: How often is "Occasionally", "Often", and "Very Often?". *Research in Higher Education, 17*, 267–281.

Pascarella, E. T., Seifert, T. A., & Blaich, C. (2009). How effective are the NSSE benchmarks in predicting important educational outcomes? *Change, 42*(1), 16–22.

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

Pike, G. R. (2006a). The convergent and discriminant validity of NSSE scalelet scores. *Journal of College Student Development, 47*, 550–563.

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

Pike, G. R. (2011). Using college students' self-reported learning outcomes in scholarly research. In S. Herzog & N. A. Bowman (Eds.), *Validity and limitations of college student self-report data* (new directions for institutional researcher series, no. 150, pp. 41–58). San Francisco: Jossey-Bass.

Pike, G. R., Kuh, G. D., & McCormick, A. C. (2011). An investigation of the contingent relationships between learning community participation and student engagement. *Research in Higher Education, 52*, 300–322.

Pike, G. R., Smart, J. C., Kuh, G. D., & Hayek, J. C. (2006). Educational expenditures and student engagement: When does money matter? *Research in Higher Education, 47*, 847–872.

Porter, S. R. (2011). Do college student surveys have any validity? *Review of Higher Education, 35*, 45–76.

Porter, S. R., Rumann, C., & Pontius, J. (2011). The validity of student engagement survey questions: Can we accurately measure academic challenge? In S. Herzog & N. A. Bowman (Eds.), *Validity and limitations of college student self-report data* (new directions for institutional research series, no. 150, pp. 87–98). San Francisco: Jossey-Bass.

Rummel, R. J. (1970). *Applied factor analysis*. Evanston: Northwestern University Press.

Ryan, J. F. (2004). The relationship between institutional expenditures and degree attainment at baccalaureate colleges. *Research in Higher Education, 45*, 97–114.

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

StataCorp (2007). *Stata 10 user's guide*. College Station, TX: StataCorp.

Toulmin, S. (1958). *The uses of argument*. Cambridge: Cambridge University Press.

Tyler, R. W. (1932). *Service studies in higher education*. Columbus: Bureau of Educational Research, Ohio State University.

Wänke, M. (2002). Conversational norms and the interpretation of vague quantifiers. *Applied Cognitive Psychology, 16*, 301–307.

Wright, D. B., Gaskell, G. D., & O'Muircheartaigh, C. A. (1994). How much is "quite a bit?" Mapping between numerical values and vague quantifiers. *Applied Cognitive Psychology, 8*, 479–496.