

The Construct Validity of Student Engagement: A Confirmatory Factor Analysis Approach

Steven M. LaNasa · Alberto F. Cabrera · Heather Trangsrud

Received: 13 November 2007 / Published online: 3 March 2009
© Springer Science+Business Media, LLC 2009

Abstract As institutions seek to promote student engagement on campus, the National Survey of Student Engagement (NSSE) is increasingly being used to chart progress and compare results using the Five Benchmark Scores. While recent research has begun to decompose the five benchmarks in a variety of ways; few research studies have sought to explore the underlying structure of these five benchmarks, their interdependence, and the extent to which the items do reflect those five dimensions. This study begins to address the instrument's construct validity by submitting a single, first-time freshman cohort's NSSE responses to a confirmatory factor analysis, and proposes as an alternative, eight "dimensions" of student engagement that fit this set of data slightly better and in a more useful way. Results have practical implications for institutions utilizing NSSE, but also contain conceptual implications pertaining to the application of these benchmarks.

Keywords Student engagement · Construct validity · Confirmatory factor analysis

Introduction

As institutions seek to foster student learning and support students striving to accomplish their educational goals, they are turning more and more of their attention to the National Survey of Student Engagement (NSSE) as an instrument to capture the impact of the institution on students and their educational activities. The popularity of NSSE is illustrated by the increased number of students and institutions participating in the annual survey. When first introduced in 2000, <300 colleges and universities that included slightly more than 60,000 students participated in the NSSE (Kuh 2000a). At its fifth year, 972 colleges and universities representing 844,000 students participated in the annual survey (Kuh 2005).

S. M. LaNasa (✉) · H. Trangsrud
Donnelly College, 608 N. 18th St., Kansas City, KS 66102, USA
e-mail: slanasa@donnelly.edu

A. F. Cabrera
University of Maryland, Baltimore, USA

Equally relevant is the fact that NSSE has also gained recognition as a potentially viable assessment tool. Charles Miller, Chair of the Commission on the Future of Higher Education, has suggested NSSE as a viable tool for institutional quality (Department of Education, 2006). One of the stronger endorsements came from, John Merrow, former education correspondent for the MacNeil/Lehrer News Hour on PBS, who even suggested in a recent podcast that using NSSE as a substitute to the US News and World Report rankings. More recently NSSE has promoted the posting of benchmark results in outlets such as *USA Today* to achieve a variety of objectives, one being to “demonstrate your institution’s commitment to transparency and accountability” (NSSE communications, 2007).

Clearly a great appeal of NSSE rests in its reliance on the well accepted seven principles of good practices in undergraduate education, as advanced by Chickering and Gamson in 1987. These seven practices, while elegant, rest on a decisively simple principle: students learn and develop when they participate in educationally purposeful activities with the different components of an institution (Kuh 2000; Chickering and Gamson 1987; NSSE 2002). Substantial research supports the engagement principles (e.g., Kuh and associates 1994) as valid predictors of learning in a variety of subject areas, growth in student competence, test scores performance, self-reported gains in general education, learning, and professional competencies (e.g., Cabrera et al. 2001; Colbeck 2001; Kuh et al. 1994, 2000; Pike 1995; Pascarella et al. 1996). While the value of engagement is well documented, the underlying sub-components and the items measuring them remain somewhat challenging, particularly for institutions that attempt to document impact on students’ development and learning. A variety of survey instruments have offered the potential to assess engagement (if not by name); however, no single instrument has garnered the attention, especially within the policy arena perhaps due to timing, as the National Survey of Student Engagement.

The Problem

NSSE represents an easy to administer instrument allowing institutions to monitor and track the educational experiences of their students. Additionally, this assessment provides an educationally grounded mechanism that facilitates comparisons with peer institutions. In other words, NSSE offers the potential to possibly serve as a robust substitute of the US News and World Report rankings because its focus is on activities actually associated with learning versus institutional inputs often which are more associated with prestige, history, and funding.

The NSSE instrument offers institutions item level data and summary institutional performance scores based on five benchmarks as indicators of student engagement (NSSE 2001, 2005). These benchmarks were developed based on past research in the areas of student learning and engagement. More specifically, these benchmarks address the areas of *academic challenge*, *active and collaborative learning*, *student-faculty interaction*, *supportive campus environments* and *enriching educational experiences*. Kuh (2003) stated that “the benchmarks were created with a blend of theory and empirical analysis” (p. 30). As institutions seek to improve and compare themselves to others, it is presumed that these benchmarks offer independent constructs, each representing a domain area that may warrant increased attention or represent a success story or best-practice.

In many respects the full reliance and use of NSSE results at an institutional level depends on two propositions: (1) that NSSE predicts relevant student outcomes, and (2) that NSSE has construct validity. In response to the first question, results are quite clear. The value of NSSE

in predicting student outcomes such as learning, critical thinking, and traditional measures of academic achievement has been established in a variety of institutional settings (Ewell 2002; Gellin 2003; Pascarella and Terenzini 2005; Carini et al. 2006; Kuh et al. 2000; LaNasa et al. 2007a); although it should be noted that in many cases only modest associations have been documented (Gordon et al. 2008). Recently, Pike (2006) found that NSSE scalelets were positively associated with student self reports of gains in general education and practical skills. However, these scalelets were constructed using an expanded list of variables beyond those used to construct the NSSE benchmark scores.

LaNasa et al. (2007b), found that several domains built with NSSE items were moderately associated with both student-reported satisfaction and first-year grade point average—but only when the variables were reconstituted and decomposed into additional dimensions. Likewise, Gordon et al. (2008) recently showed that for a single institution the benchmarks themselves accounted for little significant contribution to grade-point-average or retention for that institution, and that similarly Pike’s scalelets offered a modest improvement in predicting student outcomes compared to the baseline NSSE benchmark scores. Both studies identified significant challenges in using the scales derived from NSSE benchmarks—the most troublesome challenge being a high potential for collinearity at an institutional level.

Regarding the second proposition concerning the construct validity of the benchmarks, few studies have sought to specifically explore the overall fit of the NSSE benchmarks to single or multi-institutional data. Although the studies referenced above have begun to explore criterion-related validity (Carmines and Zeller 1979), Kane (2008) finds that “even if the test is initially validated using criterion or content evidence, the development of a deeper understanding of the constructs or processes accounting for test performance requires a consideration of construct related evidence” (p. 20). Furthermore, recent analyses have suggested the need to decompose the five-dimensions of NSSE through the incorporation of a finer breakdown of the five benchmark NSSE scales to identify additional “sub-scales” (Umbach et al. 2007; Pascarella et al. 2006; Gordon et al. 2008) or dimensions (LaNasa et al. 2007b), as well as the need to explore the convergent and discriminant validity of NSSE “scalelets” at an institutional level (Pike 2006). The findings of these studies in total seem to de-emphasize the portrayal of student engagement as a construct that is built on just five interdependent dimensions or constructs, and further suggest that it instead rests on a more expansive, complex, and interdependent set of constructs (Bean 2005). While each of these studies rely on finer decomposition of the five original NSSE benchmark, none of them addressed the issue of construct validity for the five NSSE dimensions themselves. Therefore, in light of these recent studies, the primary goal of our research was to explore the fit of various models of student engagement using one institution’s NSSE data.

Research Questions and Protocol

Two primary research questions guided our study that sought to assess the fit for the five-construct model of student engagement as described by the NSSE “benchmarks.” These questions included:

1. Does the NSSE five benchmark model fit for a single, urban institution?
2. If not, is there a factor structure model that better captures the student engagement with this institution?

In answering the first research question, we followed a construct validity approach in substantiating the presumed dimensionality of NSSE (Loevinger's 1957; Messick 1989; Kane 2008). As described above, the notion of engagement captures a complex set of interactions among students and the institutions they attend. Cronbach and Meehl (1953) stipulate that "construct validity must be investigated whenever no criterion or universe of content is accepted as entirely adequate to define the quality to be measured" (p. 282). As noted by Pike (2006), construct validity calls for a systematic examination of the extent to which an item relates to other observable variables based in theory (Carmines and Zeller 1979). Simply put, items that measure the same construct should correlate strongly among themselves—convergent validity—while displaying low correlations with those items indexing different constructs—discriminant validity (e.g., Kline 2005; Long 1983).

Figure 1 displays the initial model under consideration. Following Kuh (2000a, b), our model presumed that five dimensions account for the inter-correlations underlying the 37 items comprising the NSSE benchmarks (see description below regarding item elimination). The model also presumed that the five dimensions are inter-correlated while being defined by unique items. In testing this model, we relied on the structural equation modeling approach to construct validity (e.g., Kline 2005). This approach involved: (a) testing for a set of hypothesized number of factors, (b) assessing the extent to which the constructs are interdependent; and, (c) examining whether the items load significantly in the construct they purport to measure. Evidence against the conceptual model would be provided by poor indicators of fit, high correlations among the constructs, the presence of cross-loadings of the items in relation to constructs it is not supposed to measure, also referred as 'factorially complex' items (Cattell 1978), as well as, substantive amounts of error in the variance of the items.

In order to answer the second research question, following Pike's (2006) recommendations to explore "scalelets," we sought to examine a factor structure that better described student engagement for a Midwestern, urban institution and purposefully assessed the validity of those newly proposed dimensions. Pike suggested that institutions should consider developing scalelets, or measures that have both reliability and construct validity in order to support institutional assessment and improvement, and rely not just on the dependability of items under consideration. Pike's approach was particularly relevant for the present study because his approach is critical in rigorously reviewing the validity of scores that may serve as a potential proxy for institutional quality and accountability. Because "the objective of using scalelets in outcomes assessment is to make judgments about educational quality" (p. 552), this question was viewed as especially critical in this institution's context.

Pike's (2006) work relied heavily on Loevinger's (1957) categories of questions to define construct validity, including substantive, structural, and external components. Whereas Pike focused his examination on the external component, this study has been limited to the structural component that examines relationships among the items that comprise student engagement.

Research Design

Sample

Findings for this study were based on analyses of the NSSE results for first-time freshmen at a public, doctoral, research-intensive university in the Midwest. Approximately 60% of

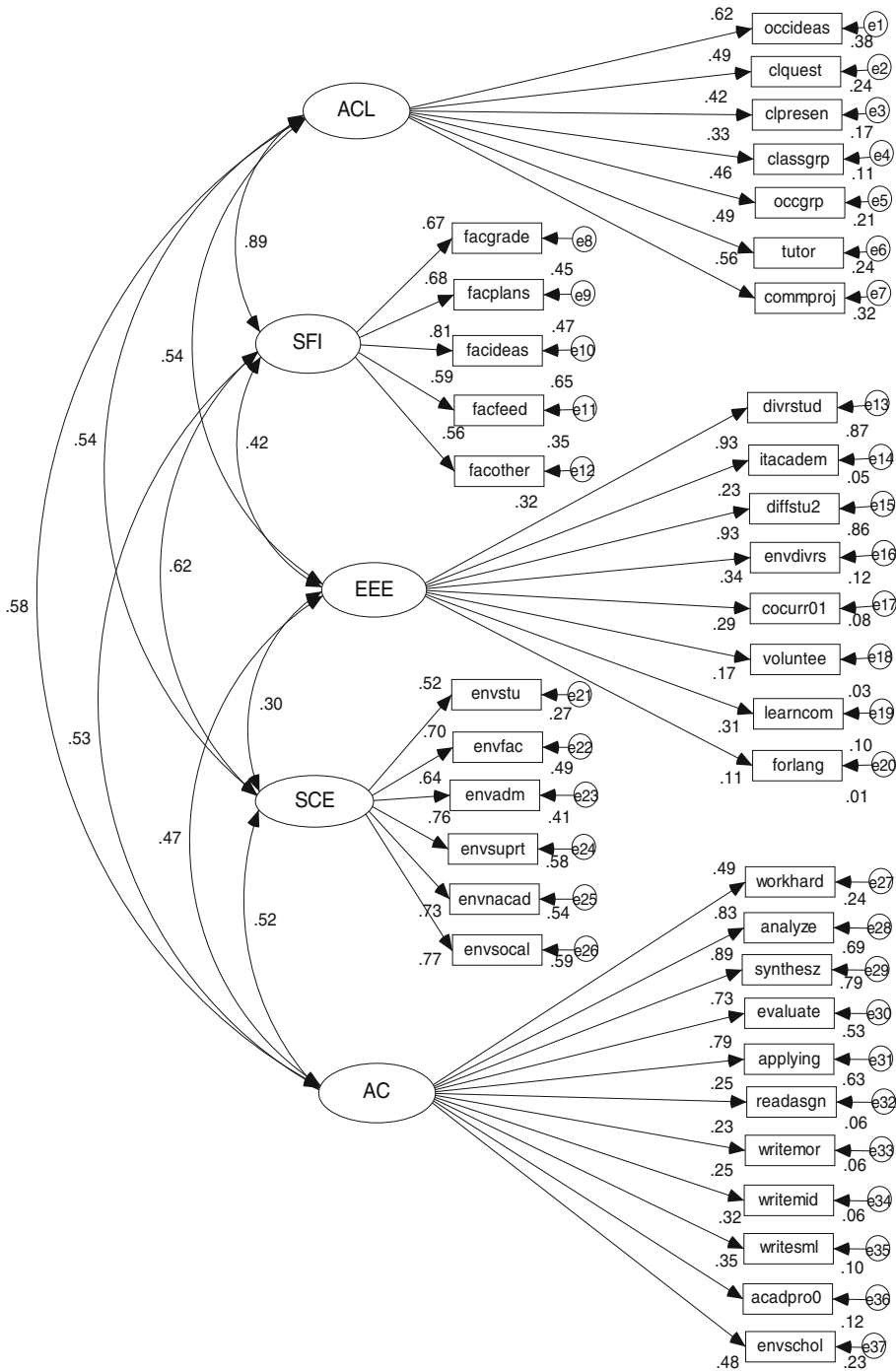


Fig. 1 Five factor model of student engagement

the first-time freshmen were female, 23% were underrepresented minorities, and 40% lived on-campus during their first year at the university. The data for this study was limited to responses of first-time freshmen from the Spring 2004 NSSE survey administrations, which were completed through the on-line web survey format. The total number of first-time freshman that completed the NSSE survey in 2004 was 375, which was a response rate of 48% and provided a representative sample of the first-time freshman populations. The Midwestern institution examined in the present study has participated in NSSE since 2000 and has expanded its use of NSSE results to foster campus-wide conversation regarding student and faculty engagement, improvement of assessment on campus, retention and success initiatives, the first-year experience, as well as identifying factors associated with graduation. For these reasons reliable measures or scales are viewed as critical.

Measures

Relying upon the NSSE benchmark calculation syntax (NSSE 2005), the researchers identified the 42 items that constitute the five NSSE benchmarks as the starting point for this analysis. A series of exploratory data analyses indicated that one NSSE indicator of Level of Academic Challenge was highly skewed; eighty percent of the students reported that they were not writing papers of 20 pages or more (*writmor*). Additionally, it also became clear that several items used to produce the benchmark scores asked students about their participation and/or their intentions to participate in various activities. On several items, almost 70% of students reported that they “planned to” participate in various activities or experiences. When these items were re-coded using NSSE syntax to create dichotomous items reflecting actual participation versus everything else, four additional items were also found to be highly skewed toward non-participation, including *seniorrx* (98.7%), *indstudy* (97.9%), *studyabr* (96.5%), *intern* (92.5%), and *research* (94.8%). Inclusion of these items appeared particularly questionable for the population of first-year students for whom many of these activities were not yet appropriate (i.e., senior year experience, internships or study abroad), as well as because of the simple fact that there was almost no variance in the items; therefore these five items were dropped from the analysis. A descriptive summary of the remaining items is found in Table 1.

Data analyses

LISREL version 8.8 (Jorskog and Sorbom 2006) was used in testing the confirmatory factor analyses model. Since NSSE has a combination of items with anchors ranging from 4 to 7, PRELIS was used in estimating the correct correlation polyserial correlation matrix as well as the corresponding asymptotic covariance matrix. Next, the weighted least square solution (WLS) was used in estimating all confirmatory factor models. By default, PRELIS uses listwise selection of cases resulting in an effective sample of 352 cases. It is important to note that the WLS employs both the polyserial and asymptotic covariance matrices in all estimation procedures. The use of this method was grounded on the scale of measurement underlying NSSE items. As noted by Joreskog and Sorbom (2006), the WLS produces correct standard errors and χ^2 values when departures of normality are present, resulting from handling non-continuous variables as is the case of NSSE items (see also Kline 2005). In estimating the confirmatory factor model (CFA), the variance of the latent constructs was set to one. This method allowed the researchers to freely estimate the loadings of the items in the corresponding construct. This approach had the additional benefit of estimating the correlation among the constructs in a direct manner, while easing the estimation of the

Table 1 Variables used in analysis (means and standard deviations)

Item	Mean	Standard deviation
1. clquest: Asked questions and contributed to class discussion.	2.67	0.89
2. clpresen: Made a class presentation.	2.22	0.79
3. classgrp: Worked with other students during class.	2.31	0.80
4. occgrp: Worked with classmates outside of class on projects.	2.37	0.88
5. tutor: Tutored other students.	1.80	0.94
6. commproj: Participated in community-based project for course credit.	1.56	0.88
7. oocideas: Discussed ideas from readings/class with others outside of class.	2.71	0.91
8. facgrade: Discussed grades/assignments with instructor.	2.57	0.87
9. facplans: Talked about career plans with faculty member.	2.06	0.92
10. facideas: Discussed ideas from readings/class with faculty.	1.77	0.85
11. facfeed: Received prompt feedback from faculty on performance.	2.56	0.86
12. facother: Worked with faculty on activities not related to class.	1.55	0.86
13. itacadem: Used electronic medium to discuss/complete assignment.	2.50	1.08
14. diverstud: Had serious conversations with students from a different race.	2.88	1.01
15. diffstud2: Had serious conversation with students from other different types of backgrounds.	2.88	0.98
16. occur01: Time spent on co-curricular activities.	1.99	1.52
17. envdivrs: Encouraged contact with students from different backgrounds.	2.72	0.99
18. analyze: Coursework emphasized analyzing content.	2.59	0.86
19. synthesz: Coursework emphasized synthesizing content.	3.14	0.83
20. evaluate: Coursework emphasized making judgments about content.	2.93	0.91
21. applying: Coursework emphasized applying content.	2.82	0.92
22. readasn: Amount of assigned reading.	3.04	0.88
23. workhard: Worked harder than you thought you could.....	2.59	0.86
24. writemid: Number of papers between 5 and 19 pages long.	2.32	0.89
25. writesml: Number of papers under five pages long.	2.93	1.12
26. acadpr01: Hours spent preparing for class.	4.25	1.87
27. envschol: Amount of time spent studying.	3.17	0.75
28. envstu: Quality of relationships with fellow students.	5.55	1.31
29. envfac: Quality of relationships with faculty.	5.15	1.34
30. envadm: Quality of relationships with admin personnel.	4.85	1.60
31. envsupt: Campus provides support needed to succeed in school.	2.90	0.83
32. envnacad: Campus helps cope with non-academic responsibilities.	1.99	0.92
33. envsocial: Campus provides support needed to thrive socially.	2.18	0.90
34. writmor: Number of papers 20 pages or more	1.24	0.606
35. volunteer: Community service or volunteer work	0.344	0.476
36. learncom: Participate in a learning community or other formal group	0.184	0.388
37: forlang: Foreign language coursework	0.243	0.429

amount of error in each item, which in itself constituted an additional indicator of goodness of the model.

Several indicators of fit were used in judging the goodness of fit of the model, which included the Non-Normed Fit Index (NNFI), also known as Tucker–Lewis Index (TLI), the

Comparative Fit Index (CFI), and the Standardized Root Mean Square Residual (RMSEA). A series of simulation studies by Hu and Bentler (1999) suggested that a cutoff value close to 0.06 for RMSEA and a cutoff value close to 0.95 are supportive of a good fit of the model in relation to the data. In addition to these indicators, the χ^2/df was used in order to lessen the sensitivity of the χ^2 test to sample size. As a rule of thumb, χ^2/df values of 3.0 or less signify a good fit of the model (Kline 2005). The use of several indicators follows Bollen's (1989) and Joreskog's (1993) recommendation of examining the extent to which the pattern of indicators is supportive of the model rather than relying on a single indicator of fit. When comparing among alternative confirmatory factor models, changes in χ^2 were computed to assess the extent to which an alternative model fitted the data better (see Kline 2005). When the models were not nested, the Akaike's Information Criterion (AIC) index was used. Low AIC values lend support of the non-nested model (see Kline 2005).

Results

Confirmatory Factor Analysis of the Five NSSE Benchmark Model

The first confirmatory factor model was used to examine the construct validity of the five-factor structure with the remaining 37 NSSE items. Consistent with Kuh's (2000) benchmarks, the model hypothesized five separate factors, those being: (1) level of academic challenge (AC), (2) student–faculty interaction (SFI), (3) active and collaborative learning (ACL), (4) enriching educational experiences (EEE), and (5) supportive campus environment (SCE). Analysis procedures tested several models beginning with the null model representing five independent constructs. Results, summarized in Table 2, depict a model that is approaching a reasonable fit of the data, with most indicators exceeding their criterion values slightly. Next, other variations were tested and compared to the null model, including versions with five interdependent constructs (Model 2), four interdependent constructs (Model 3), and five interdependent constructs with measurement error (Model 4).

Model 1 was rejected in favor of Model 2, establishing the interdependency of these data ($P < .001$). Model 3, which assumes that four constructs better capture the data because of a relatively high correlation observed between Active and Collaborative Learning and Student Faculty Interaction, did not vary significantly ($P < .223$) indicating that either model can be a plausible representation of the data. Ultimately, the researchers

Table 2 Goodness-of-fit for the original 5-Factor model and associated variations

Model	χ^2	df	χ^2/df	TLI/ NNFI	CFI	RMSEA	AIC	Changes to goodness of fit			
								$\Delta \chi^2$	P-value	Δdf	ΔCFI
1. Five independent constructs	2126.5	629	3.38	0.89	0.90	0.082	2274.5	–	–	–	–
2. Five interdependent constructs	1651.5	619	2.7	0.92	0.93	0.069	1819.5	475	0.001	10	0.03
3. Four interdependent constructs	1657.2	623	2.7	0.92	0.93	0.069	1817.2	5.7	0.223	4	0
4. Five interdependent constructs with measurement errors	1304.9	610	2.1	0.95	0.95	0.057	1490.9	346.6	0.001	9	0.02

tested the five interdependent construct model that was measured with error (Model 4) and this model was significant and deemed a best overall fit of the data. The pattern of goodness of fit indicators provided reasonable support for the hypothesized five factor structure (see Table 2), and the resulting measurement model obtained the following: $\chi^2/df > 2.1$, TLI/NNFI and CFI < 0.95 , and an RMSEA > 0.057 for this institution's data. While these model fit statistics are reasonably acceptable, item level inspection revealed measurement problems ranging from small to severe (see Table 3) and inter-correlations (see Fig. 1) that raise concern with the 5-benchmark interdependent model.

At the extreme lies the case of the Active and Collaborative Learning benchmark. None of its indicators had loadings of 0.70 or higher, suggesting that most of the variance for the seven indicators remains unexplained. The range of unexplained variance ranged from 62% to 89%. Reliability is an important consideration that has, to this point, been unaddressed. The reliability of the scale sets the upper bound for the relationship between predictor and criterion measures (Nunally and Bernstein 1994). If treated as a scale, the reliability of this benchmark would be somewhat low ($r_{xx} = 0.64$).

Moderate measurement problems were also reported by the benchmark Enriching Educational Experiences. On the one hand, two indicators were excellent manifestations of the construct with loadings greater than 0.90. Six indicators, however, showed little correlation with the benchmark. For these items, most of the variance was not captured by the construct. If treated as a scale, this benchmark would have reported a low reliability of 0.59.

Only two indicators of the benchmark Student-Faculty Interactions had loadings close to 0.70, meaning that about only half of their variance was explained by the construct they are supposed to measure. What defined this benchmark the most was discussing ideas and readings with a faculty member (*facideas*). The loading for this item was 0.81; meaning that 65% of the variance in the item was accounted by this benchmark. Evidently, the strong indicators compensated for the relative weakness of the remaining items. If treated as a scale, the reliability for this benchmark is moderately high ($r_{xx} = 0.74$).

Of the eleven indicators of Academic Challenge, only four had considerable correlations with the construct, ranging from 0.73 to 0.89. For six indicators, the variance was explained by factors other than the construct they were supposed to index. It is important to note that the reliability of the scale is moderately high ($r_{xx} = 0.75$). Evidently, the strong indicators of this benchmark compensated for the relatively low indicators. Of the five NSSE benchmarks under examination, Supportive Campus Environment appeared to be the construct best measured. Five out of the six indicators had loadings of 0.70 or higher with this benchmark. If treated as a scale, the reliability of this benchmark was relatively high ($r_{xx} = 0.79$).

The hypothesis of relative independence among the five NSSE benchmarks was not supported by these analyses because at least one of the five benchmarks was highly correlated with another (see Fig. 1). The structural correlation between Active and Collaborative Learning and Student-Faculty Interactions was just under 0.90. Moderate structural correlations were also found between Supportive Campus Environment and Student-Faculty Interactions and Academic Challenge and Active & Collaborative Learning (0.59). Only Supportive Campus Environment and Enriching Educational Experiences were relatively independent of another ($\phi = 0.29$).

Pedhazur (1982) argued that correlations among variables of 0.70 signify a high degree of multicollinearity to the point of suggesting that the two variables may be alternative measures of the same concept. In a similar vein, Tetrick, Thacker, and Fields (1989) indicated that a high degree of correlation among the constructs provides evidence that the two subcomponents may actually be representative of a single factor. As mentioned

Table 3 Loadings and variance accounted for in the five NSSE Benchmark model

Benchmark	Measure	Loading	Variance		Reliability of the scale
			Explained	Error	
Active and collaborative learning (ACL)	clquest	0.49	0.24	0.76	0.639
	clpresen	0.42	0.17	0.83	
	classgrp	0.33	0.11	0.89	
	ocgrp	0.46	0.21	0.79	
	tutor	0.49	0.24	0.76	
	commproj	0.55	0.30	0.70	
	oocideas	0.62	0.38	0.62	
Student-faculty interactions (SFI)	facgrade	0.67	0.45	0.55	0.738
	facplans	0.68	0.47	0.53	
	facideas	0.81	0.65	0.35	
	facfeed	0.60	0.36	0.64	
	facother	0.56	0.32	0.68	
Enriching educational experiences (EEE)	divrstud	0.93	0.87	0.13	0.590
	itacadem	0.23	0.05	0.95	
	diffstu2	0.93	0.86	0.14	
	envdivrs	0.34	0.12	0.88	
	cocurr01	0.21	0.08	0.92	
	volunteer	0.017	0.03	0.97	
	learncom	0.31	0.10	0.90	
	forlang	0.11	0.01	0.99	
Academic challenge (AC)	workhard	0.49	0.24	0.76	0.748
	analyze	0.83	0.69	0.31	
	synthesz	0.89	0.79	0.21	
	evaluate	0.73	0.53	0.47	
	applying	0.79	0.63	0.37	
	readasgn	0.25	0.06	0.92	
	writmor	0.23	0.06	0.94	
	writemid	0.25	0.06	0.94	
	writesml	0.32	0.10	0.90	
	acadpr01	0.35	0.12	0.88	
Supportive campus environment (SCE)	envschol	0.48	0.23	0.77	0.787
	envstu	0.52	0.27	0.73	
	envfac	0.70	0.49	0.51	
	envadm	0.64	0.41	0.59	
	envsuprt	0.76	0.58	0.42	
	envnacad	0.73	0.54	0.46	
	envsocial	0.77	0.59	0.41	

previously, a four factor confirmatory model was tested, merging the indicators of Active & Collaborative Learning and Student–Faculty Interactions into a single construct. No support was found for the four-factor benchmark model (see Table 2). The χ^2 increased slightly indicating a slightly worse fit; however, this change was not significant

($\Delta\chi^2 = 5.7$, $P = 0.223$). TLI/NNFI and CFI fit indexes were below the thresholds recommended by Hu and Bentler (1999), but the RMSEA was essentially unchanged from the five benchmark model.

Exploratory and Confirmatory Factor Analysis: An Alternative Model

The previous analysis did not fully support the five benchmark model as a definitive explanation of this institution's NSSE data. Poor item loadings and the high correlation among two of the latent constructs led the researchers to conduct an exploratory factor analyses that would attempt to better identify the dimensions explaining the data for the institution under consideration. This strategy was consistent with the exploratory factor analyses literature and with Pike's (2006) advice to identify substantive scalelets that are meaningful for the institution.

An exploratory factor analysis of the 37 NSSE items, using principal components with varimax rotation, yielded a nine-factor solution. This solution explained 61% of the variance in the correlation matrix. After examining the main loadings comprising each factor and eliminating four items that exhibited particularly low loadings of <0.34 (*writmor*, *forlang*, *volunteer*, and *learncom*), the nine dimensions were labeled as (1) Learning Strategies (Learning), (2) Academic Interaction (AcInter), (3) Institutional Emphasis (Emphasis), (4) Co-curricular Activity (Co-Curr), (5) Diverse Interactions (Diverse), (6) Effort, (7) Overall Relationships (Relations), (8) Workload, and (9) Working Collaborative In-class. These dimensions were named following a process of item inspection and analysis, as well as in conjunction with relevant student engagement literature. Next, exploratory results were submitted to a series of confirmatory factor analyses tests similar the ones used for assessing the construct validity of the 5-benchmark model.

The confirmatory factor analyses provided support for the nine factor model as shown by the pattern of goodness of fit indexes (see Table 4). The chi-square in relation to its degrees of freedom was slightly above 3. Both TLI/NNFI and CFI indices were above the threshold of 0.95. Moreover, the RMSEA value was slightly below the 0.05 threshold. The χ^2 however was significant, which is not uncommon when handling small sample sizes as the one in the present study. A close inspection of the measurement model indicated that one dimension, Working Collaboratively In-Class, was substantially weaker than the others because the items were conceptually unclear in terms of their relationship to one another and was limited to just two items. Additionally problematic was the fact that the strongest item of the pair, which focused on working with other students in class (*classgrp*), reported a loading of 0.41, signifying that more than 80% of the variance in the item was not accounted for by the factor. Consequently, after considering these problems, the ninth dimension along with its corresponding two items was dropped from the analyses.

An eight factor model fit the data better than the nine factor model as evidenced by the significant change in χ^2 ($\Delta\chi^2 = 1444$, $P < 0.001$). All measures of goodness of fit were within the acceptable ranges and the researcher selected this model as being more "parsimonious" (Kline 2005) than a nine factor solution. An examination of the pattern of structural correlations revealed a strong correlation of 0.70 between Effort and Learning Strategies, and between Academic Integration and Overall Relationships (see Fig. 2). Consequently, two alternative seven factor models were tested and compared against the eight factor model. Neither of the alternative seven-factor models provided a better fit to the data as compared to the eight-factor solution. The change in chi-square value between these models was not significant; thus in lieu of this finding, we retained the eight-factor model for further study.

Table 4 Goodness-of-fit for alternatively specified models

Model	χ^2	df	χ^2/df	TLI/ NNFI	CFI	RMSEA	AIC	Changes to goodness of fit			
								$\Delta \chi^2$	P-value	Δ df	Δ CFI
1. 9 interdependent constructs	1393.7	428	3.26	0.97	0.97	0.047	1593.72				
2. 8 interdependent constructs dropping <i>itacadem</i> and <i>classgrp</i> from the model	1248.9	377	3.31	0.97	0.97	0.047	1424.9	144.8	<0.001	51	0
3. 7 interdependent constructs merging 'learning strategies' and 'effort' into a single construct	1329.57	384	3.46	0.97	0.97	0.05	1491.57	80.67	<0.001	7	0
4. 7 interdependent constructs merging 'Institutional emphasis' and 'relationships' into a single construct	1376.17	384	3.58	0.97	0.97	0.52	1538.17	127.27	<0.001	1	0

A close inspection of the eight-factor solution revealed substantive problems with three dimensions (see Table 5). The average amount of error in Co-curricular activities, Workload, and Effort was 58%, 62% and 71%, respectively. Few of the items comprising Co-curricular and Workload had highly acceptable loadings. These two problems explained, in part, why each of these three scales report poor reliabilities. The overall amount of error associated with the items and the fact that only one benchmark had an item with a high loading (see *facother* in Co-curr), called into question these three measures for this one institution. Because Pike has argued that institutions should consider establishing institutionally specific “scalelets” for use in assessment efforts, these did not appear to be prime candidates, and were instead ones that, if these analyses are replicated elsewhere, should be avoided by institutions or they should seek to refine items that better capture this construct.

Limitations

This study is strictly limited in that it only presents analyses of a single first-time freshman cohort at an urban 4-year institution. These results may not hold for all institutional types and settings, and may not be representative of students as they mature and develop through the course of their enrollment. These results do, however, suggest that more work is needed to understand the component parts of student engagement. Engagement as a construct conveys a myriad of processes, cognitive tasks, expectations, interactions, environment, satisfaction, and even perceived successes (Bean 2005). As suggested by his Conceptual Model of College Student Engagement, “the relationships between the variables in the model will need to be specified in sub-models that include or exclude the feedback cycles. There are far too many variables in the model to be included in a single estimation of all the possible relationships indicated” (Bean 2005 p. 26). Although he was discussing his

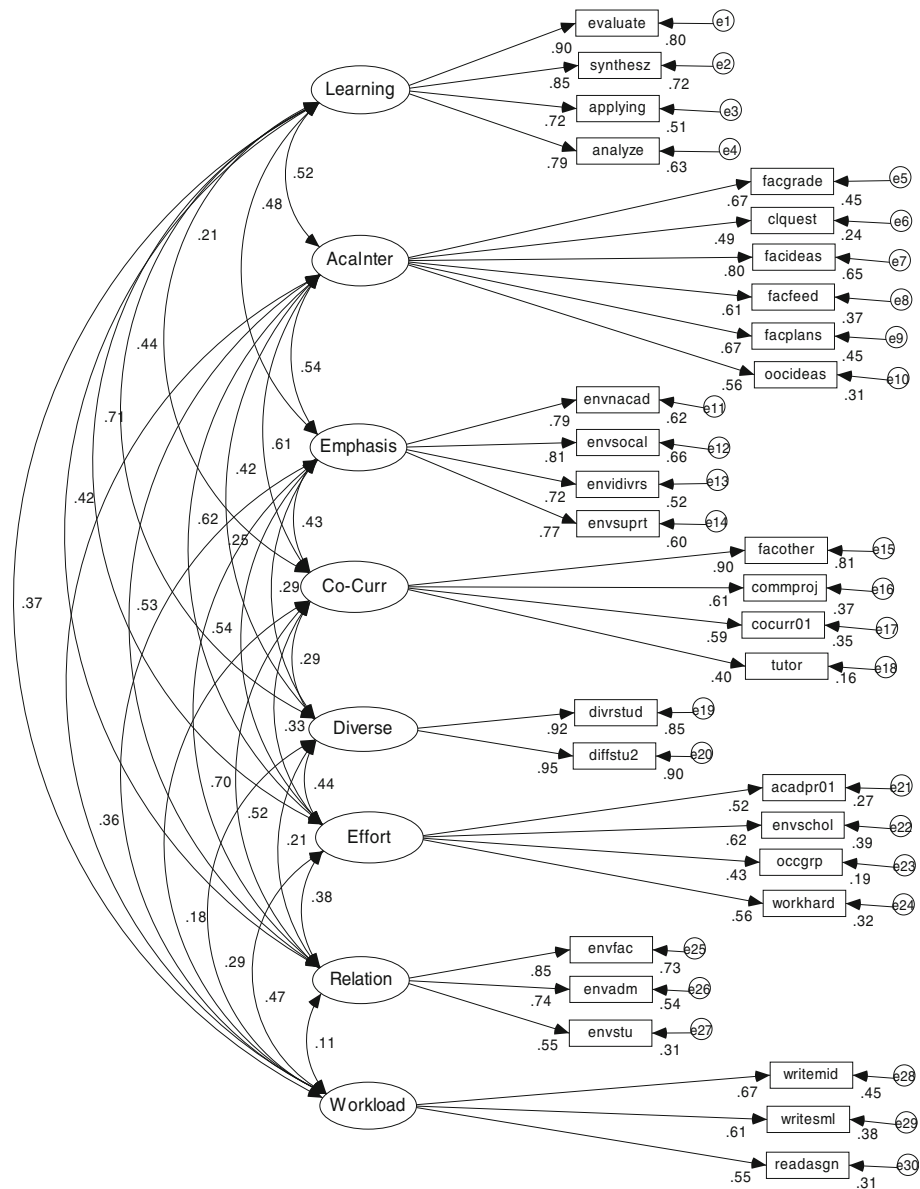


Fig. 2 Eight factor model of student engagement

more holistic model of the engagement process, his insight may be applicable to the summary of student engagement through five benchmarks that alone, based on our results, failed to capture the independence of several underlying constructs.

This study represented an examination of only a small portion of the complex construct that is student engagement, and was also likely limited in the fact that it only accounts for measurement of that complex process in just a single year. Future research must replicate this approach and examine the extent to which these underlying constructs are present in

Table 5 Loadings and variance accounted for by the eight dimension solution

Dimensions	Measure	Loading	Variance		Reliability of the scale		
			Explained	Error			
Learning strategies	synthesz	0.90	0.80	0.20	0.846		
	analyze	0.85	0.72	0.28			
	evaluate	0.72	0.51	0.49			
	applying	0.79	0.63	0.37			
Academic integration	facgrade	0.67	0.45	0.55	0.758		
	clquest	0.49	0.24	0.76			
	facideas	0.80	0.65	0.35			
	facfeed	0.61	0.37	0.63			
	facplans	0.67	0.45	0.55			
	oocideas	0.56	0.31	0.69			
	Institutional emphasis	envnacad	0.79	0.62		0.38	0.810
		envsocial	0.81	0.66		0.34	
envdivrs		0.72	0.52	0.48			
envsuprt		0.77	0.60	0.40			
Co-curricular activity	facother	0.90	0.81	0.19	0.582		
	commproj	0.61	0.37	0.63			
	cocurr01	0.59	0.35	0.65			
	tutor	0.40	0.16	0.84			
Diverse interactions	divrstud	0.92	0.85	0.15	0.888		
	diffstu2	0.95	0.90	0.10			
Effort	acadpr01	0.52	0.27	0.73	0.527		
	envschol	0.62	0.39	0.61			
	occgrp	0.43	0.19	0.81			
	workhard	0.56	0.32	0.68			
Overall relationships	envfac	0.85	0.73	0.27	0.701		
	envadm	0.74	0.54	0.46			
	envstu	0.55	0.31	0.69			
Workload	writemid	0.67	0.45	0.55	0.586		
	writesml	0.61	0.38	0.62			
	readasgn	0.55	0.31	0.69			

the data of institutions of varying types, sizes, and missions. Obviously additional research is needed to fully understand the component structure of student engagement. Future research, as called for recently by Pascarella (2006), must seek to replicate these findings across and within institutional settings.

Discussion

These results suggested that student engagement is comprised of several independent factors, in many respects somewhat separate and distinct from the five benchmark descriptions utilized by the NSSE researchers as the primary means to share results and

compare institutions. The 8-factor model was substantially improved across a range of fit statistics. Although these results were not meant to suggest that there is no meaning in the original five factors, it does suggest that a more comprehensive depiction may be required to facilitate institutional change and understanding (Pike 2006). For example, the five-factor structure subsumed several components that have been isolated in these analyses. This study's factor decomposition was fairly consistent with other recent research using NSSE data that identifies the need to further isolate components beyond the five benchmarks (i.e., Pascarella et al. 2006; Gordon et al. 2008; LaNasa et al. 2007a). Interestingly, however, especially where single institutional data is used, there may be slight variations to the structures, perhaps suggesting that engagement may “look” different at various institutions. If this is the case, the use of NSSE benchmarks for institutional comparisons deserves further attention.

To further highlight this finding, it is useful to consider the simple comparison of a benchmark “score” using the items used to produce a NSSE benchmark versus a score comprised of only those items that presented high loadings and low error in these analyses. *Enriching educational experiences* provides an excellent example. In our five factor model EEE is comprised of eight items (see description on page 4 regarding the removal of four items), but only two of those items have loadings in excess of 0.70. Using these data and the NSSE benchmark syntax the score for this scale is 46.13. But if only those items are used that possess high loadings, the score would be 62.62. This example makes clear the important relationship of the loadings on the overall validity of the “score.” Additionally, problematic from an institutional perspective is the consideration of what exactly the scores mean. In other words, does a 46.13 indicate that the institution must take steps to provide more enriching experiences on campus, or is that score a function of the high amount of error present in the score? While not easy to answer, the point is made clear that institutions are strongly encouraged to conduct a thorough analysis of NSSE data to determine what engagement looks like on their campus, and our findings further support Pike's (2006) contention for the role of “scalelets.”

As institutions grapple with the challenge of promoting increased learning and face heightened calls to document an institution's “value-added,” they are faced with an additional challenge of focusing their attention. For example, an institution that scores below average on any benchmark score may not necessarily concentrate increased efforts in the correct domains to bring about change. Does a lower than desired *enriching educational experiences* benchmark score suggest that increased diversity and exposure to diverse perspectives is needed or do students at the institution need to use computing resources more heavily or should the institution provide greater access to co-curricular activities? The benchmarks alone do not permit the easy answer of this question.

Additionally, the five-factor approach assumes that dimensions will be equivalent across institutional settings for comparative purposes, and these findings, as well as other recent studies, have called that assumption into question. While it seems reasonable to expect some smoothing of the data through a more expansive sample, the error documented in Tables 3 and 5 for both the baseline and improved models suggests that it will be important to further explore the structural model undergirding benchmark scores.

Implications for Practice

Despite these limitations, described above, or perhaps because of these limitations, this study further highlights the need for institutions to validate their results and explore the

potential need to construct items most appropriate to their institutional data, as suggested by Pike (2006). Additionally, these results suggested that institutions should closely inspect the individual items that comprise their benchmark scores. Tables 3 and 5 document that in both the baseline model and the revised, improved models that several constructs, although possessing reasonable to high reliabilities or internal consistency (Chronbach's alpha), possessed significant item-level error in their overall composition. Keeping in mind that this study used just a single institution's data, the results above suggested the strong potential for items exhibiting high levels of error to mask, or overpower, items that may actually be contributing to student engagement. Of course the converse can also be true, such that an institution might miss an area that deserves institutional focus.

The strongest implication is thus a cautionary one: It is incumbent on institutions to fully explore their own data, especially when using the data for comparative purposes. To fully understand NSSE results, item level inspection is required. An assessment of the predictive relationships of those items on desired student outcomes is also justified, but these results suggest that the benchmarks themselves are probably not sufficient. As aptly stated by Gordon et al. (2008), "...we encourage other institutions to conduct similar analyses. Through such investigations, institutions can get beyond the basic numbers to more thoroughly specify the nature of engagement within the college environment" (p. 38). The benchmarks provided an excellent jumping-off point, but for the institution to fully take advantage of the premise of student engagement, that institution must be willing to closely scrutinize the data and determine what engagement looks like for them.

Conclusion

The National Survey of Student Engagement is without a doubt a critical tool available to institutions attempting to continuously improve and examine their institutional environments and assess the relationships among what their students are doing, and what they are getting out of college. On a national scale, as the push for accountability continues to affect post-secondary education, the NSSE is being advanced as a potential tool to document institutional contributions. Because the instrument focuses on the central components of what educational institutions are focused on—namely student learning—it is not surprising that this is the case. A shift in focus to what institutions do versus what they have appears reasonable, and has long been called for by many within higher education. In some respects, this should be viewed as a positive development for higher education, although understandably threatening to some.

It is quite clear that engagement and "interactions with agents of socialization" provide a robust mechanism for examining the ways that institutional environments interact to affect student outcomes (Pascarella, 1985; LaNasa et al. 2007a). But there appears to be a difference between isolating the effects of beneficial practices or activities on a desired student outcome within an institutional context and isolating those practices for comparative purposes. Each rests on a shared set of methodological assumptions, but the development of a scale or score for comparative purposes assumes heightened expectations—especially related to the construct validity issues raised by this study. These results, if tested and replicated elsewhere as we encourage others to attempt, should provide additional evidence on the assumptions that undergird the interpretation and application of institutional comparisons and evaluation using these data.

As this discussion continues, however, based on these results, it appears critical that researchers and institutions alike attempt to further refine and assess the extent to which

student engagement as a construct is made up of five component parts. Although student engagement has been advanced as a “deceptively simple” concept, these results suggest that capturing and summarizing it in discrete terms seems to be proving less so.

References

- Bean, J. P. (November, 2005). A conceptual model of college student engagement. Paper presented at the association for the study of higher education annual meeting, Philadelphia, PA.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Cabrera, A. F., Colbeck, C. L., & Terenzini, P. T. (2001). Developing performance indicators for assessing classroom teaching practices and student learning. *Research in Higher Education*, 42(3), 327–352.
- 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.
- Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Newbury Park, CA: SAGE Publications.
- Cattell, R. B. (1978). *The scientific use of factor analysis in behavioral and life sciences*. New York: Plenum Press.
- Chickering, A., & Gamson, Z. (1987). Seven principles of good practice in undergraduate education. *AAHE Bulletin*, 39, 3–7.
- Colbeck, C. L., Cabrera, A. F., & Terenzini, P. T. (2001). Learning professional confidence: Linking teaching practices, students’ self-perceptions, and gender. *Review of Higher Education*, 24(2), 173–191.
- Cronbach, L., & Meehl, P. (1953). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302.
- Ewell, P. T. (2002). *An analysis of relationships between NSSE and selected student learning outcomes measures for seniors attending public institutions in South Dakota*. Boulder, CO: National Center for Higher Education Management Systems.
- 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.
- Gordon, J., Ludlum, J., & Hoey, J. J. (2008). Validating NSSE against student outcomes: Are they related? *Research in Higher Education*, 49(1), 19–39.
- Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1–55.
- Joreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294–316). London: Sage Publications.
- Joreskog, K. G., & Sorbom, D. (2006). *LISREL 8.8: Interactive LISREL: Technical support*. Mooresville, IN: Scientific Software.
- Kane, M. T. (2008). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.
- Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.
- Kuh, G. D. (2000a). *The NSSE 2000 report: National benchmarks for effective educational practice*. Bloomington, IN: Center for Postsecondary Research and Planning.
- Kuh, G. D. (2000b). *The national survey of student engagement: The college student report*. Bloomington, IN: Indiana University Center for Postsecondary Research and Planning.
- Kuh, G. D. (2003). What we’re learning about student engagement from NSSE: Benchmarks for effective educational practices. *Change*, 35(2), 24–32.
- Kuh, G. D. (2005). *Exploring different dimensions of student engagement: 2005 annual survey results*. Bloomington, IN: Center for Postsecondary Research and Planning.
- Kuh, G. D., Douglas, K. B., Lund, J. P., & Ramin-Gyurnek, J. (1994). *Student learning outside the classroom: Transcending artificial boundaries*. Washington, DC: The George Washington University, Graduate School of Education and Human Development.
- 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.
- LaNasa, S., Cabrera, A., Trangsrud, H., & Alleman, N. (2007a, November 9). Engagement as a proxy for learning: Testing Pascarella’s “model of engagement using NSSE items. Paper presented at the annual meeting of the association for the study of higher education, Louisville, KY.

- LaNasa, S., Olson, E., & Alleman, N. (2007b). The impact of on-campus student growth on first-year student engagement and success. *Research in Higher Education*, 40(8), 941–966.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3, 635–694.
- Long, J. S. (1983). *Confirmatory factor analysis: A preface to LISREL*. Beverly Hills, CA: Sage.
- Messick, S. (1989). Validation. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education/Macmillan.
- National Survey of Student Engagement (NSSE). (2001). *Improving the college experience: National benchmarks for effective educational practice*. Bloomington, IN: Indiana University Center for Postsecondary Research and Planning.
- National Survey of Student Engagement (NSSE). (2002). *From promise to progress: How colleges and universities are using student engagement results to improve college quality*. Bloomington, IN: Indiana University Center for Postsecondary Research and Planning.
- National Survey of Student Engagement (NSSE) (2005). Retrieved April 23, 2006, from <http://www.indiana.edu/~nsse/html/research>
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: MacGraw-Hill.
- Pascarella, E. (1985). College environmental influences on learning and cognitive development: A critical review and synthesis. In J. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 1). New York: Agathon.
- Pascarella, E. (2006). How college affects students: Ten directions for future research. *Journal of College Student Development*, 47(5), 508–520.
- Pascarella, E., Cruce, T., Umbach, P., Wolniak, G., Kuh, G., Carini, R., et al. (2006). Institutional selectivity and good practices in undergraduate education: How strong is the link? *Journal of Higher Education*, 77(2), 251–285.
- Pascarella, E., Edison, M., Nora, A., Hagedorn, L., & Braxton, J. (1996). Effects of teacher organization/preparation and teacher skill/clarity on general cognitive skills in college. *Journal of College Student Development*, 37, 7–19.
- Pascarella, E., & Terenzini, P. (2005). *How college affects students: A third decade of research*. San Francisco, CA: Jossey-Bass.
- Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and prediction* (2nd ed.). New York: Holt, Rinehart & Winston.
- Pike, G. (1995). The relationships between self-reports of college experiences and achievement test scores. *Research in Higher Education*, 36, 1–22.
- Pike, G. R. (2006). The convergent and discriminant validity of NSSE scalelet scores. *Journal of College Student Development*, 47(5), 550–563.
- Umbach, P. D., Kinzie, J., Thomas, A. D., Palmer, M. M., & Kuh, G. D. (2007). Women students at co-educational and women's colleges: How do their experiences compare? *Journal of College Student Development*, 48(2), 145–165.