How Sound Is NSSE?  
Investigating the Psychometric Properties of NSSE at a Public, Research-Extensive Institution

Corbin M. Campbell and Alberto F. Cabrera

Appraising the quality of American higher education has been dominated by the U.S. News & World Report rankings for decades (O’Meara, 2007). The popularity of the U.S. News & World Report rankings’ annual publication (2.2 million copies sold per year) attest to the importance of the rankings to both institutions and the public (Ehrenberg, 2003; Pike, 2004). Additionally, the reliance on rankings is not only present among such external actors to higher education as policymakers, parents, and prospective students (Volkwein & Sweitzer, 2006) but is also influential among such internal constituents as faculty and college administrators (Bastedo & Bowman, 2011; Meredith, 2004; O’Meara, 2007; O’Meara & Bloomgarden, 2010).

The unprecedented focus on U.S. News rankings has had some unintended consequences in the administration of colleges and universities. Institutions...
have used a number of techniques to increase their ranking, including changing admission practices, faculty reward systems, and resource allocation across departmental units (Morphew & Baker, 2004; O’Meara, 2007). Moreover, the focus of *U.S. News* on entering student characteristics (e.g., SAT scores) has led to appraising institutional quality based on student characteristics rather than the value-added that colleges and universities bestow on their students (Meredith, 2004; O’Meara, 2007; Pike, 2004; Volkwein & Sweitzer, 2006). Russell Edgerton, President of the Pew Foundation, articulated the unintended consequences of the *U.S. News* rankings in a 1999 *Chronicle of Higher Education* article. Edgerton’s warning still rings true after a decade:

> Unless we develop measures of quality where colleges can actually provide evidence of their contribution to student learning, then this whole system [of ranking colleges] turns on resources and reputation, and reinforces the elitism of higher education. (qtd. in Gose, 1999, p. A65)

In the midst of calls for changing the rankings, the public and policy arenas were seeking valid tools for assessing institutional quality (Gose, 1999). The National Survey of Student Engagement (NSSE) emerged as an alternative paradigm in institutional accountability: measuring the extent to which the students are engaged in high-impact practices as a way to assess institutional effectiveness (Kuh, 2003, 2009). A panel of experts, gathered by the Pew Foundation, was tasked with developing a survey that would measure student engagement under the well-accepted theory that more engagement means more learning (Chickering & Gamson, 1987; Kuh, Douglas, Lund, & Ramin-Gyurnek, 1994; NSSE, 2008). This survey had the potential to advance our understanding of the role of various student experiences (e.g., experiences with faculty, rigorous coursework, involvement in student organizations) in collegiate outcomes (such as persistence and learning).

The panel of experts developed the NSSE instrument using the theoretical framework of student engagement, driven by decades of research on college impact. The research found that students’ engagement with their institutions influences such important student outcomes as learning and persistence (Chickering & Gamson, 1987; Kuh, 2003; Kuh et al., 1994; NSSE, 2008). The five NSSE benchmarks were created out of the NSSE survey items using a combination of theory (specific engaging practices that seem to have the most impact on student outcomes) and exploratory factor analysis (Pike, Kuh, McCormick, Ethington, & Smart, 2011). The resulting benchmarks are intended to underscore five well-defined, different, though interrelated, constructs of undergraduate student engagement with the institution. They are presented as applicable to all types of four-year colleges and universities irrespective of their mission, Carnegie classification, location and type of students served. The five Benchmarks of Effective Educational Practice are:
The benchmarks are measured on a score of 0–100 in order “to facilitate comparisons across time, as well as between individual institutions and types of institutions” (NSSE, 2009).

The NSSE five Benchmarks of Effective Educational Practice reflect the two sides of the engagement equation: what the student does to become involved, and what the institution does to create meaningful engagement experiences. The Level of Academic Challenge benchmark ascertains the rigor of coursework in terms of academic effort and higher order thinking, such as time spent preparing for class and coursework emphasis on analysis (Pike et al., 2011). The Active and Collaborative Learning benchmark assesses whether students are asked to reflect on and apply their learning and to work with other students—for example, working with peers on projects, asking questions in class, or making class presentations. The Student-Faculty Interaction benchmark assesses students’ contact with faculty both in and out of class. The Enriching Educational Experiences benchmark covers a variety of educationally purposeful learning activities, such as study abroad experiences, conversations with diverse others, and research with faculty. Finally, the Supportive Campus Environment benchmark taps student relationships with faculty, administrators, and students and institutional support of student success. A list of the benchmarks and corresponding items is presented in the Appendix.

NSSE’s new emphasis on learning emerged as an innovative paradigm, yielding nation-wide support of the NSSE instrument (Pascarella, Seifert, & Blaich, 2008). A major endorsement of NSSE came in 2006 when Charles Miller, Chair of the Commission on the Future of Higher Education, suggested using the NSSE benchmarks as a viable tool for assessing educational quality (Miller & Malandra, 2006). Additionally, NSSE encouraged institutions to post their benchmark results in outlets such as USA Today. On the USA Today site, students and parents can see the results of the five benchmarks and how the institution compares to colleges and universities with a similar Carnegie classification.

Following the public calls for accountability and the strong endorsement of NSSE as an assessment tool, institutions with varying institutional types, locations, and missions began to use NSSE as a way to benchmark their progress and satisfy external constituents. In 1999, 56 colleges participated (NSSE, 2010a). In 2009, nearly 1,400 different colleges and universities had joined. In 2009, 443 institutions agreed to post their NSSE results to USA
Today (up from 257 in 2007). Furthermore, institutions have used the benchmark scores internally to make substantial changes in policies, to highlight effective practices, and to compare with peer institutions.

While the items making up the NSSE have strong theoretical grounding, little work has been done to investigate the construct validity of the five NSSE benchmarks themselves and the extent to which they predict relevant student outcomes (Carle, Jaffe, Vaughan, & Eder, 2009; Gordon, Ludlum, & Hoey, 2008; LaNasa, Cabrera, & Tangsrud, 2009; Pascarella, Seifert & Blaich, 2008; Pike et al., 2011; Porter, 2009). Researchers at NSSE have conducted certain analyses on the construct validity of the benchmarks including Cronbach’s alpha, stability correlations across time, and correlations across institutional types and student groups. However, to date, NSSE has not posted results on its website that use confirmatory factor analysis, item response theory, or other forms of more sophisticated techniques that are the most accepted methods for determining construct validity (Brown, 2006).

Of particular concern is the lack of research on whether the NSSE benchmarks prove reliable and valid on an institutional level (LaNasa, Cabrera, & Tangsrud, 2009; Gordon, Ludlum, & Hoey, 2008). A few studies have investigated the reliability and validity of other facets of the NSSE that are unrelated (or minimally related) to the five benchmarks (see, e.g., Carle et al., 2009; Kuh, Cruce, Shoup, Kinzie, & Gonyea, 2008; Pike, 2006). Even fewer studies examine the psychometric properties of the benchmarks themselves at an institutional level. Such a dearth of research is perplexing considering the vast use of the benchmarks by institutions in comparing across time and to other institutions/groups of institutions. On the other hand, the two studies that investigated the benchmarks at a single institution have not produced strong results confirming the internal and predictive validity of NSSE benchmarks (LaNasa, Cabrera, & Tangsrud, 2009; Gordon, Ludlum, & Hoey, 2008). Those studies found that (a) the construct validity of certain benchmarks was either marginal or poor, (b) the benchmarks did not appear to be strongly associated with important student outcomes, like GPA, and (c) the benchmarks were highly intercorrelated: They appear not to measure distinct domains of student engagement.

The paradigm shift away from *U.S. News & World Report* rankings toward measuring effectiveness through engagement is no doubt an improvement. However, as more institutions use NSSE as a tool for assessing engagement and effectiveness, the need to validate NSSE becomes increasingly critical. If the NSSE benchmarks are a valid measure of student engagement, they should be predictive of student learning across a variety of institutional types and student populations (i.e., have predictive validity). Additionally, a strong measure of institutional effectiveness must have construct validity, which in the case of NSSE, would be evidence that the five benchmarks measure five
distinct, yet intercorrelated, domains of engagement. Evidence of predictive and construct validity gathered from single- and multi-institutional samples would support the claim that the NSSE benchmarks are a valid method for ascertaining institutional effectiveness.

This study seeks to contribute to the body of research on the NSSE’s five benchmarks of effective educational practice by examining their validity and reliability at a large, public, research-extensive institution in a mid-Atlantic state. In this context, this study seeks to answer three research questions pertaining to the construct and predictive validity of the NSSE-five benchmark model:

- Are there five separate, stable benchmarks that appraise engagement?
- Do they apply to a single, large, public, research institution?
- Do they predict cumulative GPA?

**Review of the Literature**

The NSSE benchmarks were developed based on theory that links student engagement to key collegiate outcomes, such as learning and development (Kuh, 2009; NSSE, 2008; Pascarella, Seifert, & Blaich, 2008; Pike et al., 2011). In this respect, NSSE rests on a set of well-known and -respected seven practices in undergraduate education formulated by Chickering and Gamson (1987). These practices underscore a decisively simple principle: that students learn and develop when they participate in educationally purposeful activities with different aspects of an institution (Astin, 1993; Chickering & Gamson, 1987; Kuh, 2000a).

Substantial research supports the engagement practices as valid predictors of learning and development. Engagement has been positively linked to a wide range of student outcomes such as persistence (DeSousa & Kuh, 1996; Kuh et al., 2008), leadership development (Pascarella, Seifert, & Blaich, 2008; Posner, 2004), identity development (Harper, Carini, Bridges, & Hayek, 2004; Hu & Kuh, 2003), moral development (Pascarella, Seifert, & Blaich, 2008), academic performance or GPA (Carini, Kuh, & Klein, 2006; Gordon et al., 2008; Kuh et al., 2008), and critical thinking skills (Anaya, 1996; Pascarella, Seifert, & Blaich, 2008; Pike, 2000). Furthermore, student engagement is influenced by a number of factors in the collegiate environment, such as students’ involvement on campus (Astin, 1999; Kuh, Hu, & Vesper, 2000) and the frequency of interaction between students and faculty (Pascarella & Terenzini, 2005) and between students and staff (Flowers, 2003; Kuh, 2009). Precollege experiences also seem to matter in engagement. For example, students’ experiences with diverse peers prior to attending college have been found to influence collegiate engagement (Locks, Hurtado, Bowman, & Oseguera, 2008).
Predictive Validity of NSSE Benchmarks

The principle of predictive validity calls for evidence of the extent to which a test or scale predicts subsequent outcomes or behaviors that are conceptually linked to the predictor measures they are supposed to index (Krathwohl, 2004, Vogt, 2005). We identified three studies that examined the predictive validity of NSSE benchmarks. Two of them drew from multi-institutional samples (Pascarella, Seifert, & Blaich, 2008; Carini, Kuh, & Klein, 2006) and one study from a single institution (Gordon et al., 2008). Overall, research on the predictive validity of NSSE has yielded mixed results.

Pascarella, Seifert, and Blaich (2008) used data from the Wabash National Study of Liberal Arts Education to investigate the predictive validity of NSSE benchmarks on six liberal arts outcomes: critical thinking, effective reasoning and problem solving, moral character, leadership, and personal well-being. The study is compelling for two reasons. First, it uses a pre/post design, with controls for precollege abilities. Second, it explores a wide variety of college outcomes, including previously validated measures such as the Collegiate Assessment of Academic Proficiency (CAAP). The researchers conducted analyses at both the individual and institutional levels. We focus on results at the individual level since averages for only 19 institutions were used in the institutional level analyses. Pascarella, Seifert, and Blaich excluded the EEE benchmark due to its low reliability.

They also calculated upper-bound (without controls of precollege characteristics or the other NSSE benchmarks) and lower-bound estimates (with controls) of the associations between NSSE and the liberal arts outcomes. At least one of the four NSSE benchmarks was significantly and positively associated with all of the outcomes when considering the upper-bound estimates. However, when controls were introduced to account for correlations among NSSE benchmarks and precollege characteristics, many associations became non-significant. For example, the Supportive Faculty Interaction benchmark displayed a significant and negative lower-bound estimate with only one outcome, critical thinking, out of the 15 outcomes under examination. Notably, the effect size of associations that remained significant in the lower-bound estimates ranged from a negative and weak association (-.049) to small associations (.15); the median association was also rather small (.082). While the longitudinal nature of this study is promising, it is based on NSSE data from liberal arts institutions only, and further research on the predictive validity for research institutions is warranted.

Another multi-institutional study by Carini, Kuh, and Klein (2006) examined the association between the NSSE benchmarks and a few different measures of academic performance. The NSSE researchers included 1,058 first-year and senior students at 14 institutions. They reported positive but weak associations with certain NSSE benchmarks and academic achievement.
outcomes. However, Carini and colleagues did not account for intercorrelations among the NSSE benchmarks, as Pascarella, Seifert, and Blaich (2008) had done.

By contrast, Gordon, Ludlum, and Hoey (2008) focused on an individual research institution in assessing the predictive validity of NSSE benchmarks. They used multiple years of freshmen and senior data at Georgia Tech to investigate whether the NSSE benchmarks were associated with several outcomes, including freshmen retention, GPA, pursuit of graduate education, and employment outcome at time of graduation. They found that three benchmarks had minimal associations with GPA for freshmen (LAC, ACL, and EEE). Of the three benchmarks, Enriching Educational Experiences had a negative association. For seniors, only Supportive Campus Environment was significant. None of them had effect sizes larger than .3. Additionally, only one benchmark (SCE) had an association with freshmen retention. Student-Faculty Interaction was related to further education. None of the benchmarks was significantly related to employment. This study is helpful in uncovering the possible associations with multiple important outcomes. However, this study did not examine the predictive validity of each of the five benchmarks simultaneously while accounting for the degree of intercorrelation among the five benchmarks, an approach for which Structural Equation Modeling is ideally suited (Byrne, 2006; Hancock & Mueller, 2009).

**Construct Validity of NSSE**

The principle of construct validity asks for evidence of the extent to which items or scales measure the construct they purport to represent (Bollen, 1989; Brown, 2006; Byrne, 2006; Cronbach & Meehl, 1953; Kline, 2005). Porter (2009) reviewed the literature investigating the validity and reliability of the NSSE benchmarks using several methods (e.g., response bias and internal structure analyses). He found that the NSSE failed to meet validity or reliability standards. For example, he reported that the presumed structure of the five dimensions of engagement (i.e., NSSE benchmarks) had not been replicated and that reliability values consistently failed to meet basic standards (i.e., alpha levels at or above .7). Regarding the individual items within the benchmarks, he found that students were not self-reporting accurate responses and that there was ambiguous wording. He also noted that there was not sufficient evidence to support the claim that the NSSE benchmarks were associated with student outcomes.

A few scholars have investigated the construct validity of NSSE. LaNasa, Cabrera, and Tängsrud (2009) used confirmatory factor analysis to investigate the construct validity of NSSE at a single research institution. Their findings indicated that the model fit was reasonable but found several item-level measurement problems. Loadings of many items were very low, particularly
in the Active and Collaborative Learning benchmark. Moreover, they noted a high degree of overlap between the Student-Faculty Interaction and the Active and Collaborative Learning benchmarks; these two benchmarks had a large correlation of .89. For two of the five benchmarks, the reliability index was below the recommended .7 threshold. While the confirmatory approach in this study was useful in understanding the construct validity of NSSE, the data were from a single institution. Evidently, the study needs replication to determine whether the measurement problems are more widespread or isolated to this one institution.

Some research has departed from using the five benchmarks in documenting the construct and predictive validity of NSSE by reconstructing NSSE into sub-scales or scalelets (e.g., Gordon, Ludlum, & Hoey, 2008; LaNasa, Olson, & Alleman, 2007; Pike, 2006). Pike (2006) used 2001 NSSE senior data to create 12 scalelets that measured engagement and reported that some NSSE scalelets were positively associated with student self-reports of gains in general education and practical skills. Gordon, Ludlum, and Hoey (2008) later used Pike’s scalelets and found a marginal improvement over the NSSE benchmarks in terms of predictive validity but found poor reliability measures: All scalelets had alpha values less than .7. LaNasa, Olson, and Alleman (2007) found that NSSE items grouped together into eight dimensions of student engagement. Several dimensions were moderately associated with both student satisfaction and first-year grade point average. Carle et al. (2009) used freshmen and senior data from a single institution to investigate the construct validity of three alternative engagement constructs. This study was particularly useful because of the sophistication of modeling techniques (Item–Response Theory) and confirmatory factor approaches. Even George Kuh, founder of NSSE, has been using alternative engagement measures. Kuh et al. (2008) employed three separate measures from NSSE: time spent in curricular activities, time spent studying, and a “global measure of engagement” made up of 19 educational activities to determine whether engagement was associated with persistence and grades. While this line of research is informative of the applicability of NSSE to examine relevant outcomes, it is unclear why these studies would focus on alternative scales instead of researching the actual benchmarks. The NSSE five benchmarks are the scales most often used to appraise institutional effectiveness.

NSSE’s Psychometric Portfolio

Outside of peer-reviewed publications, NSSE created a “psychometric portfolio” on its website to showcase research it conducted to ensure that the survey adequately represents measures of engagement and that it is psychometrically sound (NSSE, 2010b). According to this psychometric portfolio, “As part of NSSE’s commitment to transparency as well as continuous
improvement, we routinely assess the quality of our survey and resulting
data, and we embrace our responsibility to share the results with the higher
education community” (NSSE, 2010b). The psychometric portfolio includes
research briefs on studies of reliability, validity, and other quality indicators.

The validity section of the portfolio includes studies of seven types of
validity: response process, content, construct, concurrent, predictive, known
groups, and consequential. Of particular relevance to the present study are the
NSSE portfolio briefs on construct and predictive validity. In the construct
validity brief, NSSE reports the results of a second order confirmatory factor
analysis that substantiates a construct called “deep learning,” which has three
sub-scales: higher-order, integrative, and reflective learning. It is pertinent to
note that there is no portfolio brief on the construct validity of the five NSSE
benchmarks of effective educational practices, which are the primary scales
used by institutions to report NSSE results. The brief on predictive validity
includes a study that found a positive relationship between four of the five
NSSE benchmarks (all but SFI) and persistence/credits earned.

The reliability section of the psychometric portfolio on the NSSE website
offers research briefs on three aspects of reliability: internal consistency,
temporal stability, and equivalence (NSSE, 2010b). For internal consistency,
they report Cronbach’s alpha levels each year for the five benchmarks, which
ranged from .591 (EEE) to .787 (SCE) for freshmen in 2009. The NSSE website
also reports the intercorrelations among the benchmarks (mainly moderate)
and inter-item correlations within each benchmark (varies by benchmark,
but mainly low to moderate).

It is clear that researchers at NSSE have conducted a wide array of psycho-
metric studies on the properties of NSSE, but there are two glaring omissions.
First, they have not reported construct validation of the five benchmarks
of effective educational practices. Second, they cite no research examining
how well the benchmarks hold true for individual institutions. In fact, Kuh
(2009) acknowledges: “Institution-specific analysis sometimes produce[s]
factor structures different than the five benchmarks or clusters of effective
educational practices that NSSE uses to report its findings” (p. 687).

**Methods**

*Data Collection*

The focus institution contracted with NSSE to survey the population of
senior students in the spring of 2009 (N = 5,117). The NSSE survey included
42 items that comprised the five benchmarks, among other items related
to student engagement and institutional activities. (See Appendix for the
benchmark items and variable names.) The survey was administered by
NSSE via the web, yielding a response rate of 28%.
**Data Sources**

This study used 2009 NSSE data from a large, public, research-extensive university. Only non-transfer seniors were utilized, yielding an analytical sample of 1,026, or approximately 28% of the population of senior students at the time of the survey. Non-transfer seniors were chosen because they had had four years to experience and be engaged with the institution. Institutional variables such as college cumulative GPA, high school GPA, and SAT math scores were added to the dataset.

**Procedures**

First, we obtained descriptive analyses to determine assumptions related to normality of the data. We then recoded certain items to categorical or dichotomous due to low variability and non-normality. Other items were recoded to match the NSSE coding of items for creating benchmarks.

Following Kuh (2000a, 2000b, 2004), we tested a confirmatory factor model assuming that five dimensions account for the intercorrelations underlying the 42 items comprising the NSSE benchmarks. The model also presumed that the five dimensions are intercorrelated while being defined by unique items. Following recommendations from the Structural Equation Modeling literature (e.g., Bollen, 1989; Brown, 20006; Byrne, 2006; Kline, 2005), we adopted a two-step strategy in answering our three research questions. First, we used Confirmatory Factor Analysis (CFA) to answer the first two research questions, regarding whether there are five stable and distinct benchmarks. We set the variance of each latent construct associated to each benchmark to one, allowing us to ascertain the extent to which the items indeed loaded in their corresponding conceptual latent factor (Brown, 2006; Kline, 2005).

Next, we employed Structural Equation Modeling to answer our third research question regarding whether the benchmarks predict college cumulative GPA. While some scholars have debated whether GPA is useful as a sole predictor of achievement or collegiate success (Baird, 1985; Porter, 2009), there is strong evidence that college GPA is a relevant criterion of success in college. For example, it is related to graduation which is an ultimate goal of college (Cabrera, Burkum, & LaNasa, 2005). College GPA has a demonstrated relationship to admission to graduate school, and certain forms of occupational success (Allen, Robbins, Casillas, & Oh, 2008; Baird, 1985; Carini, Kuh, & Klein, 2006; Pascarella & Terenzini, 2005; Tinto, 1993). Additionally, studies that have investigated the relationship between student engagement and college success often use GPA (as a proxy for academic ability) and persistence as the outcome variables of interest. (See, for example, Kuh et al., 2008; Gordon, Ludlum, & Hoey, 2008; Carini, Kuh, and Klein 2006). Finally, GPA has been found to be one of the best predictors of persistence in college (e.g., Pascarella & Terenzini, 2005; Cabrera, Nora, & Castañeda, 1992).
The model testing the predictive validity of NSSE on college GPA included a measure of precollege academic ability that was a composite factor of high school GPA and SAT math score. High school GPA and SAT math score have been consistently cited in the literature as predictors of college academic success and persistence. (See, for example, Allen et al., 2008; Pascarella & Terenzini, 2005; Tinto, 1993).

We relied on four robust measures of fit to judge the CFA model and the SEM models. These indices include: (a) the Satorra-Bentler Maximum Likelihood estimate of chi-square ($S-B\chi^2$), (b) the Comparative Fit Index (CFI), (c) the Non-Normed Fit Index (NNFI), and (d) the Root Mean Square Error of Approximation (RMSEA). We guided our selection of goodness-of-fit values based on recommendations from the SEM literature (Byrne, 2006; Hu & Bentler, 1999). Accordingly, we sought CFI and NNFI values of 0.95 or higher to signify an excellent fit, but we also considered values greater than .90 to be appropriate. In terms of RMSEA, we judged values ranging from 0 to .05 excellent, but we also considered RMSEA values less than .08 to be suitable. In addition, we estimated 90% ($CI_{90}$) confidence intervals to check that RMSEA values did not fall beyond the cut off value of .10, signifying the rejection of the model. In contrasting alternative models underscoring NSSE benchmarks, we relied on the Satorra-Bentler corrected for non-normality chi-square test (Byrne, 2006; Hancock & Mueller, 2009).

**Limitations**

A few limitations of this study are noteworthy. First, this study is based on NSSE data from a single, large, research-extensive institution. While this study contributes to our understanding of the validity of NSSE on an institutional level, results may not be generalizable to other institutions, particularly those of other institutional types.

Second, this study investigates only GPA as a measure of student success. There is evidence that GPA is one useful measure (Allen et al., 2008; Baird, 1985; Carini, Kuh, & Klein, 2006; Pascarella & Terenzini 2005; Tinto, 1993), and cumulative GPA is often a criterion for graduation. Moreover, recent research has demonstrated that cumulative GPA is an excellent predictor of degree completion (Cabrera, Burkum, & LaNasa, 2005). Still, further studies should investigate the predictive validity of NSSE in relation to other measures of success (e.g., graduation, other measures of student learning).

A third limitation is that this study investigates only non-transfer seniors. If other populations were included, results could be different. However, non-transfer seniors likely have the best ability to respond to a survey of student engagement that appraises institutional effectiveness due to their longevity of experiences with the institution. A senior sample may be more engaged
than a freshmen sample (due to the relationship between engagement and persistence), but the NSSE benchmarks should be valid for both the most engaged and the least engaged students. Additionally, certain benchmarks include items that are most relevant to seniors, such as participation in a capstone experience or internships.

Lastly, the response rate was 28%, which is somewhat lower than the NSSE average response rate in 2009 of 36%. However, the NSSE average response rate includes all participating institutions from all institutional types. It is possible that research institutions have slightly lower response rates.

**Results**

We conducted analyses using EQS software. After running initial descriptive analyses, we determined that 10 variables should be recoded and treated as dichotomous or ordinal; the remaining 32 variables were treated as continuous.

The first EQS model revealed that the data departed from normality. The standardized Mardia value of 6.71 was higher than the threshold of 5.0. Consequently, we relied on stepwise least robust methods contained in EQS 6.1 (version 2010) in our model estimation. This method adjusts estimates due to departure of normality and can handle ordinal and continuous variables such as those present in this study (Bentler, 2005; Byrne, 2006).

**Construct Validity of NSSE**

In testing our first two research questions, we used confirmatory factor analysis (CFA). The first CFA model tested the hypothesis that NSSE captures five benchmarks or factors that are independent of one another. It became our null model, in comparison to an alternative model that tested the hypothesis that the 42 items in NSSE account for five correlated constructs or benchmarks. Table 1 reports the results of testing these two alternative representations of NSSE.

Our first model is non-tenable. The TLI/NNFI and CFI fall far below the .90 threshold. The RMSEA for the model is .085 with a tenuous 90% confidence interval (CI$_{90}$).

The second model approached NSSE as representing five interrelated benchmarks. As shown in Table 1, this model represents a significant improvement over the null model ($\Delta S-B \chi^2 = 89.6, p < .05$). While the five correlated benchmarks...
<table>
<thead>
<tr>
<th>Model</th>
<th>Robust Measures of Goodness of Fit</th>
<th>Changes in Robust Measures of Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ML $\chi^2$</td>
<td>S-B $\chi^2$</td>
</tr>
<tr>
<td>1. Five independent constructs</td>
<td>6,117.3</td>
<td>8,107.2</td>
</tr>
<tr>
<td>2. Five interdependent constructs</td>
<td>5,252.8</td>
<td>2,806.3</td>
</tr>
</tbody>
</table>
benchmarks model is better than the five uncorrelated factors model, it does not appear to be a good representation of the data. Both the NNFI (.795) and the CFI (.807) did not support the model, while the RMSEA value was acceptable (.056, CI90 = .053, .058).

A close examination of factor results reveals substantial correlations among the latent constructs representing the NSSE five benchmarks. (See Table 2.) Notable is the .864 correlation between the Active & Collaborative Learning and Student-Faculty Interactions benchmarks, suggesting that these two constructs may be indexing only one domain of student engagement. Equally interesting is the strong correlation between Enriching Educational Experiences and Active and Collaborative Learning (.749). The lowest correlation between benchmarks was between Supportive Campus Environment and Level of Academic Challenge, which was still a moderate-sized correlation (.435).

Table 3 presents the loadings corresponding to each of the five benchmarks. More than half of the NSSE items have poor loadings, well below the recommended 0.500 threshold (Kline, 2005). Eight items had loadings less than .3; six additional items had loadings less than .4; and eight additional items had loadings less than .5. This leaves only 20 out of 42 items with loadings greater than the recommended threshold. When looking at each benchmark separately, the Supportive Campus Environment and Student Faculty Interaction benchmarks have a majority of items with loadings greater than the recommended threshold. By contrast, the Level of Academic Challenge, Active and Collaborative Learning, and Enriching Educational Experiences include a majority of items with loadings less than the recommended threshold.

\[2\text{We tested an alternative model to NSSE, one in which ACL and SFI benchmarks were collapsed into a single factor. This alternative four-factor model was not supported. The corrected chi-square indicated that the alternative model significantly worsened the fit (}\Delta S-B \chi^2 = 130, p < .05). \text{Moreover, all robust measures of fit for this model were below the recommended threshold with the exception of the RMSEA.}\]
Additionally, many items appear to explain little variance in the model and have substantial error. (See Table 3.) For the Level of Academic Challenge benchmark, the average percentage error across the 11 items was 79%. The variance explained for each item in the LAC benchmark ranged from 2% to 42%, with six of eleven accounting for less than 15% of variance. For the Active and Collaborative Learning benchmark, the average percent of error across the seven items was 79%. Variance explained for each item in ACL ranged from 14% to 28%. The average percent of error across the six items in the Student-Faculty Interactions benchmark was 64%, with the variance explained ranging from 9% to 48%. For the Enriching Educational Experiences benchmark, the average percent of error across the 12 items was 84%. Variance explained for each item ranged from 1% to 31%, with less than 10% of variance explained for five out of twelve items. Of the five benchmarks, the Supportive Campus Environment benchmark had the least error in its items, with an average of 63% across loadings, with the variance explained ranging from 25% to 49%.

In total, the average percent of error across the item loadings in each benchmark ranged from 63% to 84%. Particularly noteworthy is the average percent of error in items for the Level of Academic Challenge (79%), Active and Collaborative Learning (79%), and Enriching Educational Experiences (84%) benchmarks. Additionally, certain items seem to explain almost no variance and are almost entirely error. For example, one item in the Level of Academic Challenge benchmark accounts for only 2% of variance with 98% remaining of error. One item in the Enriching Educational Experiences benchmark accounts for only 1% of variance, with 99% remaining of error.

**Predictive Validity of NSSE**

In answering the third research question, we relied on Structural Equation Modeling to estimate the predictive validity of the five NSSE benchmarks in relation to college cumulative GPA, when controlling for high school academic ability (GPA and SAT math score). Results showed that the model explained 56% of the variance in cumulative GPA, and the RMSEA value of .057 is acceptable (CI_{90} = .054, .059). However, the CFI = .78 and the NNFI = .77 are well below the .90 thresholds. As a result, when all the Goodness of Fit indices are considered, the model is not a strong representation of the data. (See Table 4.)

Moreover, the results demonstrate that there are only two individual paths that have a significant influence on college cumulative GPA (p < .05). The first is high school academic achievement (GPA and SAT math score), which was set as a control in this study. Expectedly, the effect of high school academic achievement on college cumulative GPA is quite strong (.663). By
### Table 3

**Loadings and Variance Accounted for in the Five NSSE Benchmark Model**

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Measure</th>
<th>Loading</th>
<th>Variance</th>
<th>Average % of error across loadings</th>
<th>Reliability of the Benchmark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level of</td>
<td>ACADPR01</td>
<td>.317</td>
<td>.100</td>
<td>79%</td>
<td>.661</td>
</tr>
<tr>
<td>Academic Challenge</td>
<td>WORKHARD</td>
<td>.542</td>
<td>.294</td>
<td>0.706</td>
<td></td>
</tr>
<tr>
<td></td>
<td>READASGN</td>
<td>.146</td>
<td>.021</td>
<td>0.979</td>
<td></td>
</tr>
<tr>
<td></td>
<td>WRITEMOR</td>
<td>.338</td>
<td>.114</td>
<td>0.886</td>
<td></td>
</tr>
<tr>
<td></td>
<td>WRITEMID</td>
<td>.267</td>
<td>.071</td>
<td>0.929</td>
<td></td>
</tr>
<tr>
<td></td>
<td>WRITESML</td>
<td>.205</td>
<td>.042</td>
<td>0.958</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ANALYZE</td>
<td>.611</td>
<td>.374</td>
<td>0.626</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SYNTHESZ</td>
<td>.650</td>
<td>.422</td>
<td>0.578</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EVALUATE</td>
<td>.631</td>
<td>.398</td>
<td>0.602</td>
<td></td>
</tr>
<tr>
<td></td>
<td>APPLYING</td>
<td>.623</td>
<td>.389</td>
<td>0.611</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ENV SCHOL</td>
<td>.324</td>
<td>.105</td>
<td>0.895</td>
<td></td>
</tr>
</tbody>
</table>

2. Active & Collaborative Learning

|                | CLQUEST    | .435    | .189     | 79%                                | .618                        |
|                | CLPRESEN   | .486    | .236     | 0.764                              |                             |
|                | CLASSGRP   | .379    | .144     | 0.856                              |                             |
|                | OCC GRP    | .440    | .193     | 0.807                              |                             |
|                | TUTOR      | .425    | .180     | 0.820                              |                             |
|                | COMM PROJ  | .532    | .283     | 0.717                              |                             |
|                | OCC IDEAS  | .486    | .237     | 0.763                              |                             |

3. Student-Faculty Interactions

|                | FAC GRADE  | .516    | .266     | 64%                                | .708                        |
|                | FAC PLANS  | .690    | .476     | 0.524                              |                             |
|                | FAC IDEAS  | .588    | .346     | 0.654                              |                             |
|                | FAC FEED   | .557    | .310     | 0.690                              |                             |
| FACOTHER | .579 | .335 | 0.665 |
| RESRCH04 | .299 | .089 | 0.911 |

4. Enriching Educational Experiences

| DIVRSTUD | .518 | .268 | 0.732 | 84% | .543 |
| DIFFSTU2 | .508 | .258 | 0.742 |
| ITACADEM | .448 | .200 | 0.800 |
| ENVDIVRS | .463 | .214 | 0.786 |
| INTERN04 | .394 | .155 | 0.845 |
| VOLNTR04 | .557 | .310 | 0.690 |
| LRNCOM04 | .511 | .262 | 0.738 |
| FORLNG04 | .106 | .011 | 0.989 |
| STABR04  | .192 | .037 | 0.963 |
| INDSTD04 | .282 | .079 | 0.921 |
| SNRX04RE | .268 | .072 | 0.928 |
| COCURR01 | .305 | .093 | 0.907 |

5. Supportive Campus Environment

| ENVSUPRT | .701 | .492 | 0.508 | 63% | .758 |
| ENVNACAD | .639 | .408 | 0.592 |
| ENVSOCAL | .632 | .399 | 0.601 |
| ENVSTU   | .498 | .248 | 0.752 |
| ENVFAC   | .641 | .411 | 0.589 |
| ENVADM   | .536 | .287 | 0.713 |
contrast, out of the five benchmarks, only Enriching Educational Experiences has a significant effect on cumulative GPA (p < .05), and this effect is rather strong (.653 in standardized units). The resulting model is shown in Figure 1, displaying the correlations among the five benchmarks and the paths between each benchmark and GPA.

We also ran a model that included only the benchmarks as a predictor of GPA, without high school academic achievement as a control. The model explained only 21% of the variance in cumulative GPA. In terms of goodness of fit indices, the model represents a poor fit of the data. The CFI = .79 and the NNFI = .774 are well below the .90 thresholds. However, the RMSEA value of .057 is acceptable (CI90 = .054, .059). Similar to results of the model that included precollege academic achievement, Enriching Educational Experiences was the only benchmark that had a significant effect on cumulative GPA (p < .05), and this effect is rather strong (.686 in standardized units).

**Discussion**

In 2009, Porter found that NSSE did not meet validity and reliability standards. In particular, he raised questions about the extent to which engagement items (a) have content validity, (b) prompt accurate responses from students, and (c) group together into five distinct domains as postulated by Kuh (2004). If Porter’s conclusions are correct, one would expect to find NSSE items displaying poor loadings in their corresponding conceptual latent factors, lack of support on behalf of the underlying five-dimensionality of NSSE, and poor predictive power as well.

All in all, Porter’s hypotheses seem to hold true for this study. Our confirmatory factor analyses show that the five-benchmark model does not hold for the focus institution. Of particular note was the substantial overlap between Active & Collaborative Learning with Student-Faculty Interactions
Campbell & Cabrera / How Sound Is NSSE?

95

...with a structural correlation of .864) and Enriching Educational Experiences (with a structural correlation of .749). In other words, the amount of overlap between these two pairs of benchmarks ranged from 56% to 75%, which questions the extent to which these three benchmarks appraise distinct dimensions of student engagement.

Second, each benchmark does not appear to be well appraised by the individual items, as evidenced by poor alpha reliabilities and low factor loadings. Particularly troublesome is the amount of error found in the Enriching Educational Experiences benchmark, with eight of 12 items having loadings lower than .5. If treated as a scale, the alpha reliability of this benchmark is .543, which lies well below the .70 recommended threshold (Porter, 2009). The only benchmarks that are moderately well appraised by the NSSE items are Supportive Campus Environment and Student-Faculty Interaction benchmarks with an average percent of error across loadings of about 60% and alpha reliabilities at or slightly above the recommended .70 threshold. (See Table 3.)

In terms of validity, we found that the model linking the five benchmarks with GPA represented a poor fit of the data. Two out of three indicators of
fit rejected this model. Under this questionable model, only one of the four benchmarks, Enriching Educational Experiences, is predictive of GPA. Paradoxically, this benchmark is the least reliable.

Our results seem to both contradict and replicate those of the extant literature. While our predictive validity results appear to be in sharp contrast with those based on multi-institutional samples (e.g., Carini, Kuh, & Klein, 2006; Pascarella, Seifert, & Blaich, 2008), they are in alignment with those based on samples drawn from a single institution (e.g., Gordon, Ludlum, & Hoey, 2008). For instance, Carini, Kuh, and Klein reported significant, though relatively small, partial correlations between four of the NSSE benchmarks and GPA among 1,058 undergraduate students drawn from 14 four-year colleges and universities. On the other hand, our results are consistent with those reported by Gordon, Ludlum, and Hoey (2008) and LaNasa, Cabrera, and Tangsrud (2009). Gordon, Ludlum, and Hoey investigated the predictive validity of the benchmarks at a major, doctoral-extensive institution. They reported small, if not negligible, validity coefficients for the association between Level of Academic Challenge and Active Collaborative Learning benchmarks and GPA among freshmen. Only Supportive Campus Environment was found to have a significant, though small, effect size on seniors’ GPA. Additionally, LaNasa, Cabrera, and Tangsrud relied on a sample of freshman students from a Midwestern university. Like our study, they, too, noted that the NSSE five-benchmark model was a poor representation of the data. They also observed that the NSSE benchmarks substantially overlapped among themselves.

As far as reliability is concerned, our results are quite consistent with the literature in singling out Enriching Educational Experiences to be the least reliable benchmark (LaNasa, Cabrera, & Tangsrud, 2009; Gordon, Ludlum, & Hoey, 2008; Pascarella, Seifert, & Blaich, 2008). As a matter of fact, this benchmark was deemed so unreliable that Pascarella and associates excluded it from their validation study of NSSE as a predictor of outcomes linked to liberal arts education.

**Conclusion**

Recently, Pascarella, Seifert, and Blaich (2008) judged NSSE benchmarks to be good measures of precursors of liberal arts outcomes. This conclusion, based on a solid research design that is longitudinal with powerful controls, heightens the argument on behalf of NSSE as a viable substitute of *U.S. News & World Report* rankings to appraise institutional quality at liberal arts institutions. Our study, however, raises questions about the validity of NSSE to appraise quality for research-extensive institutions. The results of this study in combination with the study by LaNasa, Cabrera, and Tangsrud (2009) suggest that, at least for two research-extensive institutions, the NSSE
benchmarks did not hold. It is also important to note that, even for the liberal arts institutions, some questions about validity remain. The Pascarella, Seifert, and Blaich (2008) study reported small effect sizes and untenable reliability scores for the EEE benchmark. They also did not investigate the construct validity of the benchmarks.

Our findings question the extent to which NSSE benchmarks are a universal tool for appraising institutional quality, and whether they predict such student outcomes as GPA. We echo Gordon, Ludlum, and Hoey’s (2008) advice to institutional researchers and policymakers. They should carefully examine the extent to which the five NSSE benchmarks are reliable and valid for their own institutional contexts before committing themselves to major organizational changes. If each of the five benchmarks does not measure a distinct dimension of engagement and includes substantial error among its items, it is difficult to inform intervention strategies that will improve undergraduates’ educational experiences. For example, if it is unclear what the EEE benchmark actually measures, the interventions could be targeting the wrong precursor of learning.

Additionally, our results, in consonance with research at other institutions, may suggest the refinement of NSSE benchmarks to be more reliable and valid measures. NSSE plans to launch an updated version of the survey in 2013, dubbed, “NSSE 2.0” (NSSE, 2011). Our research would suggest that the NSSE researchers pay special attention to the psychometric properties of the revised benchmarks at the institutional level to ensure that NSSE 2.0 reaches its full potential as an instrument to measure engagement and student learning across institutions.

The changing paradigm from U.S. News rankings to measuring student engagement is both courageous and more reflective of institutional quality. Yet if the NSSE benchmarks are not psychometrically sound, they have the potential to misinform such institutional stakeholders as prospective students, administrators, and policymakers.

**APPENDIX**

**Level of Academic Challenge (11 items)**

About how many hours do you spend in a typical seven-day week doing each of the following? (0, 1–5, 6–10, 11–15, 16–20, 21–25, 26–30, more than 30)

- **acadpr01**: Time spent preparing for class (studying, reading, writing, rehearsing, and other activities related to your academic program)

In your experience at your institution during the current school year, about how often have you done each of the following? (very often, often, sometimes, never)

- **workhard**: Worked harder than you thought you could to meet an instructor’s standards or expectations
During the current school year, about how much reading and writing have you done? (None, 1–4, 5–10, 11–20, More than 20)
readasgn: Number of assigned textbooks, books, or booklength packs of course readings
writemor: Number of written papers or reports of 20 pages or more
writemid: Number of written papers or reports between 5 and 19 pages
writesml: Number of written papers or reports fewer than 5 pages

During the current school year, how much has your coursework emphasized the following mental activities? (very much, quite a bit, some, very little)
analyze: Analyzing the basic elements of an idea, experience, or theory such as examining a particular case or situation in depth and considering its components
synthesize: Synthesizing and organizing ideas, information, or experiences into new, more complex interpretations and relationships
evaluate: Making judgments about the value of information, arguments, or methods, such as examining how others gathered and interpreted data and assessing the soundness of their conclusions
applying: Applying theories or concepts to practical problems or in new situations

To what extent does your institution emphasize each of the following? (very much, quite a bit, some, very little)
ensschol: Spending significant amounts of time studying and on academic work

Active and Collaborative Learning (7 items)
In your experience at your institution during the current school year, about how often have you done each of the following? (very often, often, sometimes, never)
clquest: Asked questions in class or contributed to class discussions
clpresen: Made a class presentation
classgrp: Worked with other students on projects during class
occgrp: Worked with classmates outside of class to prepare class assignments
tutor: Tutored or taught other students (paid or voluntary)
commprey: Participated in a community-based project (e.g., service learning) as part of a regular course
occideas: Discussed ideas from your readings or classes with others outside of class (students, family members, co-workers, etc.)

Student-Faculty Interaction (6 items)
In your experience at your institution during the current school year, about how often have you done each of the following? (very often, often, sometimes, never)
facgrade: Discussed grades or assignments with an instructor
facplans: Talked about career plans with a faculty member or advisor
facideas: Discussed ideas from your readings or classes with faculty members outside of class
facother: Worked with faculty members on activities other than coursework (committees, orientation, student life activities, etc.)
facfeed: Received prompt written or oral feedback from faculty on your academic performance
Which of the following have you done or do you plan to do before you graduate from your institution? (done, plan to do, do not plan to do, have not decided)

resrch04: Work on a research project with a faculty member outside of course or program requirement

**Enriching Educational Experiences (12 items)**

In your experience at your institution during the current school year, about how often have you done each of the following? (very often, often, sometimes, never)

diffstu2: Had serious conversations with students who are very different from you in terms of their religious beliefs, political opinions, or personal values

divrstud: Had serious conversations with students of a different race or ethnicity than your own

itacadem: Used an electronic medium (listserv, chat group, internet, instant messaging, etc.) to discuss or complete an assignment

To what extent does your institution emphasize each of the following? (very much, quite a bit, some, very little)

envdivrs: Encourages contact among students from different economic, social, and racial or ethnic backgrounds

Which of the following have you done or do you plan to do before you graduate from your institution? (done, plan to do, do not plan to do, have not decided)

intern04: Practicum, internship, field experience, co-op experience, or clinical assignment

volntr04: Community service or volunteer work

forlng04: Foreign language coursework

stdabr04: Study abroad

indstd04: Independent study or self-designed major

snrx04: Culminating senior experience (capstone course, senior project or thesis, comprehensive exam, etc.)

lnrcmm04: Participate in a learning community or some other formal program where groups of students take two or more classes together

About how many hours do you spend in a typical 7-day week doing each of the following? (0, 1–5, 6–10, 11–15, 16–20, 21–25, 26–30, more than 30)

cocurr01: Participating in co-curricular activities (organizations, campus publications, student government, fraternity or sorority, intercollegiate or intramural sports, etc.)

**Supportive Campus Environment (6 items)**

To what extent does your institution emphasize each of the following? (very much, quite a bit, some, very little)

envsuprt: Provide the support you need to help you succeed academically

envnacad: Help you cope with your non-academic responsibilities (work, family, etc.)

envsoc: Provide the support you need to thrive socially
Mark the box that best represents the quality of your relationships with people at your institution

envstu: Relationships with other students (continuum 1 = unfriendly, unsupportive, sense of alienation; 7 = friendly, supportive, sense of belonging)

envfac: Relationships with faculty members (continuum 1 = unavailable, unhelpful, unsympathetic; 7 = available, helpful, sympathetic)

envadm: Relationships with administrative personnel and offices (continuum 1 = unhelpful, inconsiderate, rigid; 7 = helpful, considerate, flexible)

REFERENCES


at the Annual Meeting of the Association for the Study of Higher Education, Jacksonville, FL.


