

To the Graduate Council:

I am submitting herewith a dissertation written by Shelley Leigh Esquivel entitled “The Factorial Validity of the National Survey of Student Engagement.” I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Educational Psychology and Research.

Schuyler Huck, Major Professor

We have read this dissertation
and recommend its acceptance:

John Lounsbury

Jennifer Richards

Gary Skolits

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

The Factorial Validity of the National Survey of Student Engagement

A Dissertation Presented for

The Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Shelley Leigh Esquivel

May 2011

Dedication

To Mario, my loving husband of over 10 years.

You wanted me to earn this degree as much as I wanted it for myself.

Without your support, I couldn't have done this.

This is for you...because of you....

To my two beautiful daughters, Lucy and Cora.

You both continue to teach me about what's truly important in life.

I love you all.

Acknowledgments

I would like to thank my dissertation committee for their support and guidance throughout this entire process: Sky Huck (chair), Gary Skolits, Jennifer Richards, and John Lounsbury. Thanks to my friends at the Office of Institutional Research and Assessment, particularly Kent Wagoner, Olivia Halic, and Elizabeth Pemberton, for talking through my ideas and guiding me along the way. To Mario, thank you for not getting offended or upset when I locked myself up in the office for hours at a time while I worked on this. I also want to thank my daughters, Lucy and Cora, for unknowingly giving me the motivation to achieve my goal of earning this degree.

Abstract

The purpose of this research was to explore the factorial validity of the National Survey of Student Engagement (NSSE), a survey widely used by institutions of higher education. Specifically, using data collected from first-year students and seniors at The University of Tennessee, Knoxville (UT), this research addressed three research questions. First, to what extent does the five-factor model of NSSE (i.e., the benchmark model) exhibit factorial validity? Second, to what extent is Pike's (2006b) scalelet model of the NSSE factorially valid? Finally, is there a model that depicts the NSSE data better than the models consisting of benchmarks or scalelets? The participants of this study were first-year ($n = 981$) and senior ($n = 944$) students at UT who completed the online version of the NSSE in the spring of 2009. Using confirmatory factor analysis, results suggested poor model fit for both the benchmark model and Pike's (2006b) scalelet model. Exploratory factor analysis with oblique rotation (Promax) resulted in a six-factor solution consisting of 27 items that accounted for approximately 39 percent of variance. The six-factor model failed, however, to exhibit sufficient model fit when confirmatory factor analysis was applied to a different data set (i.e., NSSE data collected in the spring of 2010). Overall, results suggest that much more validation research is needed for the National Survey of Student Engagement to ensure that its use among institutions of higher education is appropriate.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION	1
Purpose	2
Rationale.....	2
CHAPTER 2: LITERATURE REVIEW	5
Validity.....	5
Evidence Based on Test Content	7
Evidence Based on Response Processes.....	9
Evidence Based on Internal Structure	9
Evidence Based on Relationships Among Variables	14
Evidence Based on the Consequences of Testing	18
The National Survey of Student Engagement	19
Conceptual Framework.....	19
NSSE Items and Benchmarks.....	20
How Institutions Use NSSE	25
Reliability of NSSE Benchmarks	28
Validity of NSSE Benchmarks	29
NSSE Scalelets.....	41
Summary of Literature Review	45
CHAPTER 3: METHOD	46
Instrument.....	46
Procedure.....	46
Participants	47
Confirmatory Factor Analysis.....	47
Data Considerations	56
Sample Size and Missing Data	56
Normality and Outliers	57

NSSE Benchmarks: Is the 5-Factor Model of Student Engagement Valid?	58
Internal Consistency	58
Confirmatory Factor Analysis	58
Pike's Scaletts: Is the 12-Factor Model of Student Engagement Valid?.....	61
Internal Consistency	61
Confirmatory Factor Analysis	61
Is There a Model That Better Fits the Data for The University of Tennessee, Knoxville?	63
Factor Analysis	63
Confirmatory Factor Analysis Using 2010 Data	64
CHAPTER 4: RESULTS	66
Normality and Outliers.....	66
NSSE Benchmarks: Is the Five-Factor Model of Student Engagement Valid?	69
Internal Consistency	69
Confirmatory Factor Analysis	70
Summary of Benchmark Results	73
Pike's Scaletts: Is the 12-Factor Model of Student Engagement Valid?.....	74
Internal Consistency	74
Confirmatory Factor Analysis	75
Summary of Scalelet Results	77
Is There a Model That Better Fits the Data for The University of Tennessee, Knoxville?	78
Factor Analysis	78
Confirmatory Factor Analysis	80
Localized areas of strain	80
Examination of parameters	81
Summary of Factor Analysis and CFA of Six-Factor Model	81
Overall Summary of Results	82
CHAPTER 5: DISCUSSION.....	83
Summary of Findings	83

Discussion	85
LIST OF REFERENCES	89
APPENDIX	100
VITA	131

LIST OF TABLES

Table A.1. <i>NSSE Benchmark Items</i>	101
Table A.2. <i>Pike's Scalelet Items</i>	104
Table A.3. <i>Normality Assessment</i>	108
Table A.4. <i>Benchmarks: Internal Consistency</i>	110
Table A.5. <i>Benchmarks: Analysis of Residuals</i>	113
Table A.6. <i>Benchmarks: Modification Indices</i>	114
Table A.7. <i>Benchmarks: Parameters</i>	115
Table A.8. <i>Scalelets: Internal Consistency</i>	117
Table A.9. <i>Scalelets: Analysis of Residuals</i>	120
Table A.10. <i>Scalelets: Modification Indices</i>	121
Table A.11. <i>Scalelets: Parameters</i>	122
Table A.12. <i>Factor Analysis: Initial Pattern Matrix</i>	124
Table A.13. <i>Final Factor Analysis Solution</i>	126
Table A.14. <i>Six-Factor Model: Analysis of Residuals</i>	127
Table A.15. <i>Six-Factor Model: Modification Indices</i>	128
Table A.16. <i>Six-Factor Model: Parameters</i>	129

LIST OF FIGURES

Figure 3.1. <i>Measurement Model</i>	49
Figure 3.2. <i>The Five-Factor NSSE Benchmark Model</i>	60
Figure 3.3. <i>Pike's (2006b) Scalelet Model</i>	62
Figure 4.1. <i>Normality: Histograms</i>	67
Figure 4.2. <i>Factor Analysis: Scree Plot</i>	79

CHAPTER 1: INTRODUCTION

The National Survey of Student Engagement (NSSE) (Kuh, Hayek, Carini, Ouimet, Gonyea, & Kennedy, 2001) is a survey that is administered to first-year and senior undergraduate students at institutions of higher education nationwide. The survey, which in its 2010 edition consisted of 85 items (excluding demographic questions), “...assesses the extent to which students...are participating in educational practices that are strongly associated with high levels of learning and personal development” (Kuh, 2001a, p. 12).

Although the instrument is popular among institutions of higher education, with 603 colleges and universities in the United States administering the instrument in the spring of 2010 (NSSE, n.d.(a)), the five subscales (which NSSE refers to as *benchmarks*) comprising the NSSE have not exhibited sufficient factorial¹ validity in previous research (LaNasa, Cabrera, & Trangsrud, 2009; Swerdzewski, Miller, & Mitchell, 2007). Specifically, LaNasa et al. (2009), after determining that the five benchmarks did not fit their data well, used principal components analysis (PCA) and found that a nine-factor model provided a better fit. Using confirmatory factor analysis (CFA), Swerdzewski et al. (2007) similarly concluded that the benchmarks did not fit their university’s NSSE data. Without data to support the factorial validity of the NSSE, researchers cannot confidently measure the construct of student engagement on their campuses (Carle, Jaffee, Vaughan, & Edar, 2009).

Another problem with the benchmarks, according to Pike (2006b), is that they are too broad to be used for program improvement. For this reason, he suggested that researchers and assessment professionals group NSSE items into what he called *scalelets*, which consist of “a

¹ Factorial validity is often referred to as *structural validity*. In this study, factorial validity is defined as “the extent to which items designed to measure a particular factor (i.e., latent construct) actually do so” (Byrne, 2010, pp. 97-98).

limited number of survey questions that provide a measure of a specific aspect of the educational experiences of a group of students” (p. 181). Because Pike encouraged the use of these smaller subscales by institutional researchers “who are charged with taking NSSE results and translating them into a series of action items to improve the student experience on campus” (Pike, 2006a, p. 558), the factorial validity of his proposed NSSE scalelets must be explored.

Purpose

The purpose of this research was to see whether evidence could be garnered which supports the validity of the factorial structures of the NSSE benchmark and scalelet (Pike, 2006b) models, using data from The University of Tennessee, Knoxville (UT) population of students. Specifically, this research addressed three questions. For the population of students at The University of Tennessee, Knoxville,

1. To what extent does the five-factor model of NSSE (i.e., that which is represented by five benchmarks) exhibit factorial validity?
2. To what extent is Pike’s (2006b) model of the NSSE, comprised of scalelets, factorially valid?
3. Is there a model that depicts the NSSE data better than the models consisting of benchmarks or scalelets?

Rationale

Data obtained from administrations of the National Survey of Student Engagement has been used by institutional researchers, academic advisors, college counselors, and researchers interested in student learning and related constructs (NSSE, n.d.(a)). Important decisions, based at least partly on NSSE results, are made by administrators and others in universities and

colleges across the nation. Not only is information shared within institutions, NSSE results are shared externally. Kuh (2001b) succinctly described this by stating that, "...the most frequently mentioned external groups with which institutions expect to share their NSSE results are national accreditation agencies and professional accrediting groups, state higher education commissions, foundations, prospective students and their parents, and alumni" (p. 17). In fact, institutional funding is often tied to results from the NSSE. For instance, performance funding by the Tennessee Higher Education Commission (THEC) allows institutions to receive up to 5.45 percent of their operations budgets if they meet established standards. Student scores on the NSSE benchmarks provide the basis for 10 percent of possible points in the first and fourth year of the current cycle (2010-2015).

NSSE has partnered with *USA Today* so that individual institutions can publish their results for others to see. Such readily available information makes it easy for prospective students, as well as institutional stakeholders, to compare institutions of higher education based on NSSE scores. Some researchers have even discussed the possibility of NSSE data replacing current institutional rankings (LaNasa et al., 2009; Pascarella, Seifert, & Blaich, 2010). LaNasa et al. (2009), for instance, discussed the possibility that NSSE "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 which are more associated with prestige, history, and funding" (p. 316). Recently, the *Knoxville News Sentinel* published an article with the title, "UT students say they're less satisfied than peers," and described the NSSE items as "indicators for student retention rates and graduation rates" (Boehnke, 2010). To summarize, NSSE data are

becoming a more readily available source of information for several groups of people, including institutional researchers, prospective students, accrediting agencies, and even the media.

Surprisingly few studies (e.g., LaNasa et al., 2009; Swerdzewski et al., 2007) have explored the reliability and validity of the NSSE instrument, particularly in terms of its factorial validity. Those studies that were done are described in Chapter 2, but for now suffice it to say that *much more* research in this area is necessary before confidence can be placed on NSSE results.

CHAPTER 2: LITERATURE REVIEW

The current chapter is a review of the literature in three areas. First, a review of the literature on instrument validation is warranted. What does it mean for an instrument to “have validity”? What kinds of evidence are needed to support an assertion that a particular instrument “is valid”? Next, a discussion of the NSSE instrument follows, including a description of its conceptual framework, its items and benchmarks, and its use among institutions of higher education. This portion of the literature review is both descriptive and evaluative. The greatest proportion of this discussion focuses on the reliability and validity evidence accrued for the NSSE instrument. Using the framework provided in the first part of the literature review (i.e., the review of the validation literature), previous research is reviewed and critically evaluated. The last component of the literature review focuses on the NSSE scalelets as proposed by Pike (2006b). This portion of the literature review is relatively brief compared to the first two portions due to the small amount of research that has been done on Pike’s (2006b) scalelets.

Validity

The purpose of this investigation was to determine whether evidence can be accumulated that supports the validity of the NSSE benchmarks and scalelets (both of which are described in more detail below). Validity refers to the adequacy and appropriateness of the *inferences* drawn from a test or other measure, not the measure itself, so it is the *use* of an instrument that is validated (Kane, 1992; Koeske, 1994; Messick, 1993). According to Messick (1993), inferences can be considered hypotheses, and the validation of those inferences is akin to hypothesis testing. Kane (1992) described what he called an *argument-based approach* to validity:

A test score interpretation always involves an interpretive argument, with the test score as a premise and the statements and decisions involved in the interpretation as conclusions. The inferences in the interpretive argument depend on various assumptions, which may be more-or-less credible (p. 527).

For example, inferences about theoretical constructs (e.g., student engagement) based on test scores (e.g., the NSSE) require assumptions included in the theory defining the construct (e.g., that student engagement is related to student outcomes such as retention) (Kane, 1992).

Determining whether these assumptions are “more-or-less credible,” as Kane (1992) put it, is the goal of validation.

The many, often conflicting, terms used in the validation literature have resulted in confusion over what is meant by “validity” (Koeske, 1994). By far, the main source of confusion has been surrounding the various definitions of *construct validity* (e.g., American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA/APA/NCME], 1999; Anastasi & Urbina, 1997; Cronbach & Meehl, 1955; Messick, 1993). In the past, the purpose of construct validity was described as exploring “internal and external test structures, that is, [the] patterns of relationships among item scores or between test scores and other measures” (Messick, 1993, p. 17). According to more recent standards, construct validity encompasses all possible validity evidence (AERA/APA/NCME, 1999; Messick, 1993). In fact, the trend has been to move away from the use of the phrase *construct validity* because “all test scores are viewed as measures of some construct, so the phrase is redundant with validity” (AERA/APA/NCME, 1999, p. 174).

According to the Standards for Educational and Psychological Testing (AERA/APA/NCME, 1999), “validity is a unitary concept. It is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed purpose” (p. 11).² Sources of validity evidence for a measure include evidence based on its internal structure, as well as evidence based on the measure’s associations with other variables or constructs (Hoyle & Smith, 1994). Specifically, validity evidence refers to : (a) test content, (b) response processes, (c) internal structure, (d) relations to other variables, and (e) the consequences of testing (AERA/APA/NCME, 1999). The types of validity evidence needed generally depend on the purpose of the instrument (Benson & Clark, 1982; Messick, 1993). The focus of the current study is on validity evidence based on the internal structure of student engagement as depicted by the NSSE benchmarks and Pike’s (2006b) scalelets, but a discussion of all types of validity evidence is warranted.

Evidence Based on Test Content

Validity evidence based on test content is “obtained from an analysis of the relationships between a test’s content and the construct it is intended to measure” (AERA/APA/NCME, 1999, p. 11). To accomplish this, many things need to be explicitly articulated. First, it is necessary to identify the specific objectives of the instrument, as well as the content areas to be included in the measure (Benson & Clark, 1982). A description of the types of items and the underlying constructs they are intended to assess (and *not* meant to measure) should be explicitly stated before items are created (AERA/APA/NCME, 1999; Benson & Clark, 1982; Clark & Watson,

² Much of the literature about validity refers to tests, but in general, the same criteria are applied to other measures or instruments, such as surveys.

1995; Messick, 1995). The items should be designed to sufficiently capture the essence of each construct (Anastasi & Urbina, 1997; Cronbach & Meehl, 1955).

Content validity refers not only to the test items and their underlying constructs, but also to the formatting of the measure and the guidelines for administration and scoring it (AERA/APA/NCME, 1999). Evidence for validity based on test content “includes logical or empirical analyses of the adequacy with which the test content represents the content domain and of the relevance of the content domain to the proposed interpretation of test scores” (AERA/APA/NCME, 1999, p. 11). Expert judgment is usually necessary to establish this kind of validity evidence (AERA/APA/NCME, 1999; Benson & Clark, 1982); specifically, experts in the field should determine that the items on an instrument reflect the dimensions they are meant to measure (Benson & Clark, 1982). When experts are used to gather evidence for a measure’s content validity, several things need to be reported: the number of experts, their qualifications, the directions they were provided, and the extent to which experts agreed or disagreed (Anastasi & Urbina, 1997).

The evaluation of a measure’s content validity has at least two other implications. First, it can lead to the modification of the constructs being assessed, as well as the identification of constructs not yet thought of (Smith & McCarthy, 1995). Second, the process of gathering content validity evidence can ensure that the items in a measure are unidimensional (i.e., that they measure one construct and not several), and can lead to the addition of items that were initially excluded (Smith & McCarthy, 1995). The unidimensionality of a measure (or subscale) is one of the subjects addressed in a later section which focuses on validity evidence based on the internal structure of a measure.

Evidence Based on Response Processes

Validity evidence based on response processes concerns “the fit between the construct and the detailed nature of performance or response actually engaged in by examinees” (AERA/APA/NCME, 1999, p. 12). This type of validity evidence has more to do with *how* people are responding to a test or other instrument, such as a survey, than *what* they are responding to. In other words, it entails collecting “empirical evidence of response consistencies” (Messick, 1995, p. 745). This type of evidence is necessary for the overall validity of a measure because participants may interpret items differently even when other psychometric evidence is strong (Ouimet, Carini, Kuh, & Bunnage, 2001). These different interpretations by participants are an important source of variability that may be accounted for by the researcher simply by observing participants as they respond to the instrument (Cronbach & Meehl, 1955).

Methods used to collect validity evidence based on participant response processes are diverse and depend on the measures or instruments used, as well as the constructs being studied (AERA/APA/NCME, 1999). Common methods include questioning participants about their reasons for responding to items in a particular way, and examining the eye movements or response times of participants (AERA/APA/NCME, 1999).

Evidence Based on Internal Structure

A researcher wanting to know *what* is being measured by a particular scale might want to know more about the internal structure of the scale. In other words, what constructs are actually being measured, and how are these constructs related to one another? Also, what items on the instrument are associated with each construct? The answers to these questions can be partly addressed using one or more of the techniques described below (e.g., internal consistency).

Answering these questions is an important task, because the “definition of instrument structure is a prerequisite to subsequent instrument refinement” (Smith & McCarthy, 1995, p. 301).

As previously mentioned, the terms used in the validation literature are not always consistently applied, or even clearly defined (e.g., AERA/APA/NCME, 1999; Anastasi & Urbina, 1997; Cronbach & Meehl, 1955). In the past, some authors have referred to validity evidence for the internal structure of a measure as contributing to its *construct validity* (e.g., Benson & Clark, 1982; Cronbach & Meehl, 1955). In more recent writing, construct validity (if the phrase is used at all) tends to refer to *all* validity evidence (e.g., AERA/APA/NCME, 1999). In the review that follows, it is often the case that the author(s) were using the phrase *construct validity*, but the phenomena they were describing were at least somewhat aligned with what the Standards (AERA/APA/NCME, 1999) refer to as *validity evidence based on internal structure*. For instance, Cronbach and Meehl (1955) defined construct validation as being “involved whenever a test is to be interpreted as a measure of some attribute or quality which is not ‘operationally defined.’ The problem faced by the investigator is, ‘What constructs account for variance in test performance?’” (p. 282). This question is of interest when the trait or quality underlying the test is important, as opposed to the test behavior or the relationship between the test and some other criterion (American Psychological Association/American Educational Research Association/National Council on Measurements Used in Education, 1954). Obviously, the exact nature of the hypothesized structure of a construct dictates the specific analyses to be conducted, but the “extent to which item interrelationships bear out the presumptions of the framework” is important when assessing validity (AERA/APA/NCME, 1999, p. 13).

Internal consistency and unidimensionality. A theory that predicts unidimensionality (or homogeneity) among a group of test items (i.e., one that predicts a single underlying construct), would need evidence indicating a high internal consistency among items (AERA/APA/NCME, 1999; Cronbach & Meehl, 1955). *Internal consistency* refers to the degree of interrelatedness among items that purportedly contribute to a single construct (Cortina, 1993; Hoyle & Smith, 1994). Measures of internal consistency also provide an estimate of reliability (AERA/APA/NCME, 1999). In fact, a criticism of traditional psychometric concepts is that reliability is not usefully separated from the idea of validity when examining appropriateness of test use [because] the concept of reliability is directly related to construct meaning (validity), for example, the degree to which the measured concept should exhibit score stability, or reflect internal item homogeneity” (Barnett, Lentz, & MacMann, 2000, p. 369).

Internal consistency does not automatically imply unidimensionality (Briggs and Cheek, 1986; Cortina, 1993); rather, it is best to think of internal consistency as “a necessary but not sufficient condition for homogeneity” (Cortina, 1993, p. 100). Cortina (1993) succinctly described a common measure of internal consistency, coefficient alpha (Cronbach, 1951), as “...a function of the extent to which items in a test have high communalities and thus low uniquenesses. It is also a function of interrelatedness, although one must remember that this does not imply unidimensionality or homogeneity” (p. 100). It is quite possible to have a scale or a subscale that has a high coefficient alpha, yet also has low item intercorrelations or even multiple dimensions (Cortina, 1993). Another important point to make here is that, when evaluating coefficient alpha, one needs to account for the number of items in a scale (Cortina, 1993).

Researchers should not deem a coefficient alpha of .70, for example, as acceptable evidence of internal consistency without also demonstrating that the value of coefficient alpha is relatively invariant when there are many items in a measure (Cortina, 1993). For these reasons, other measures (e.g., interitem correlations, the mean interitem correlation) might be more useful indicators of internal consistency (Briggs & Cheek, 1986).

Factorial validity. Factor analysis may be used to explore the factors comprising an instrument (Benson & Clark, 1982; Briggs & Cheek, 1986; Cronbach & Meehl, 1955; Harrington, 2009; Koeske, 1994). Factorial validity is “the degree to which the measure of a construct conforms to the theoretical definition of the construct (c.f., Loevinger, 1957)” (Hoyle & Smith, 1994, p. 432). Koeske (1994) described factorial validity, and the role of factor analysis in determining the internal structure of constructs:

Factor analysis permits assessment of the structure of constructs, resulting in inferences of what is usually called factorial validity. These procedures may best be seen as informing content validity, but are typically presented in the literature as reflecting construct validity. In these procedures a measure is evaluated by examining its internal structure.... In the typical case, a measure is assumed to acquire validity if the set of items is found to have a factor structure corresponding to expectation (p. 56).

The last sentence of Koeske’s (1994) statement illustrates the importance of theory. It is theory that dictates what the structure of a construct (or entire instrument) should look like (Koeske, 1994). Without a clear idea of what a construct is *supposed to* look like, information regarding its structure is meaningless. Once the researcher has an idea of what a construct is supposed to look like, confirmatory factor analysis can be used to determine (a) whether a

construct is unidimensional or multidimensional in nature (i.e., the number of factors), (b) whether, and to what extent, the factors comprising an instrument are correlated, and (c) if items are related to constructs as intended during the creation of the instrument (Harrington, 2009; Hoyle & Smith, 1994).

Factor analysis is a tool that helps researchers understand what a scale is measuring (Briggs & Cheek, 1986). Briggs and Cheek (1986) described how the factors resulting from factor analysis can resolve ambiguities based on previous analyses:

[Factors] resolve the ambiguity of...low interitem correlations, they are interpretable, and they are related to other measures of various sorts in sensible ways. Not only do factors clarify how the scale works psychometrically, they reveal something crucial about the construct itself (p. 129).

Repeatedly confirming a factor structure with factor analysis by using different samples for each investigation allows one to be reasonably confident that the scale is consistently measuring the constructs of interest (Benson & Clark, 1982).

As stated earlier, the specific analyses to be performed depend on the hypothesized relationships among variables (Messick, 1993). In some cases, it does not make any sense to examine internal consistency (Bollen & Lennox, 1991). If test items in a subscale are thought to cause the construct of interest (i.e., when indicator variables are *causal indicators*), for example, the extent to which test items are interrelated does not matter (Bollen & Lennox, 1991). If, on the other hand, test items depend on the latent variable (that is, when indicator variables are *effect indicators* that are determined by the construct), measures of internal consistency make sense (Bollen & Lennox, 1991). To summarize, effect indicators associated with the same construct

should be positively correlated; the same does not hold true for causal indicators (Bollen & Lennox, 1991).

There is disagreement in the literature about the extent to which items on a measure designed to assess a single construct should be correlated (e.g., Bollen & Lennox, 1991; Briggs & Cheek, 1986; Cattell, 1965; Clark & Watson, 1995; Comrey, 1988; Smith & McCarthy, 1995). For instance, Clark and Watson (1995) have recommended that all items in a scale (or subscale, if applicable) should be moderately correlated; that is, interitem correlations should all range from .15 to .50. Their reasoning for suggesting moderate, rather than high, interitem correlations is based on the *attenuation paradox*, which states that increasing internal consistency beyond a certain point will not enhance its construct validity and in fact might result in its attenuation (Clark & Watson, 1995). In contrast, Bollen and Lennox (1991) “see no reason to choose indicators with moderate correlations over those with high correlations for effect indicators” (p. 307). Consistent with Bollen and Lennox’s (1991) conclusion, Smith and McCarthy (1995) asserted that items should be “parallel, alternative indicators of the same, underlying construct” (p. 302) to ensure that items are reflective of a unidimensional construct. A goal of theory-driven assessment is, after all, to “measure a single construct systematically, [and] the test developer ultimately is pursuing the goal of homogeneity or unidimensionality rather than internal consistency per se” (Clark & Watson, 1995, p. 316).

Evidence Based on Relationships Among Variables

The extent to which a measure is (or is not) related to other variables “addresses questions about the degree to which these relationships are consistent with the construct underlying the proposed test interpretations” (AERA/APA/NCME, 1999, p. 13). These other

variables, which are external to the measure itself, include instruments that are expected to be either the same as or different from the measure of interest (e.g., convergent and discriminant validity evidence), as well as those that the measure is expected to predict (e.g., predictive validity evidence) (AERA/APA/NCME, 1999). A discussion of these types of validity evidence follows.

Convergent and discriminant validity. Clark and Watson (1995) articulated the importance of convergent and discriminant validity evidence when they wrote "...a good theory articulates not only what a construct is, but also what it is not" (p. 311). By considering a measure's convergent and discriminant validity, one can determine a construct's boundaries and gain knowledge about what the scale does and does not measure (Clark & Watson, 1995). A definition of convergent validity was provided by Koeske (1994): "*Convergent* validation applies when an inference of accurate construct assessment is made from data demonstrating a relationship between measures of the *same* construct assessed by *different* measures" (p. 51). Importantly, the methods that are used to establish convergent validity should be different because of the likelihood of data being correlated simply because they were collected using similar methods (i.e., the problem of "shared method variance" that can result when, for instance, both instruments are self-report surveys) (Koeske, 1994). The process of collecting convergent validity evidence described by Foster and Cone (1995), although perhaps oversimplified, is helpful: "One need merely obtain scores on the measure for a group of persons and scores on an independent measure of the same latent variable and correlate them. High correlations support convergence" (p. 250).

On the other hand, when results from a measure are *not* highly correlated with measures or characteristics theoretically targeting *different* constructs, evidence for discriminant validity has been amassed (Koeske, 1994). Theory plays an important part of collecting validity evidence for both convergent and discriminate validity, without which evidence for validity is not possible (Harrington, 2009; Koeske, 1994). In other words, theory guides the researcher when making decisions about whether two hypothetically different constructs are, in the case of discriminant validity, too highly correlated to infer that the construct of interest is valid (Koeske, 1994). Another important requirement when establishing convergent and divergent validity, in addition to theory, is the validity of the target measures used (Koeske, 1994). Otherwise, correlations among measures that are inconsistent with a researcher's theory may result from the use of a poor target instrument rather than instrument invalidity.

Test-criterion relationships. “Usefulness, in any predictive sense, is not an intrinsic property of factorially derived scales...but must be demonstrated empirically” (Lanyon & Goodstein, 1997, p. 118). One of the final stages of test validation involves the acquisition of evidence pertaining to the relationships among (a) instrument constructs and (b) other variables of interest to the researcher. Such validity evidence is necessary when an instrument will be used to make predictions (Benson & Clark, 1982). According to many authors, this type of validity evidence is best collected after the demonstration of other types of validity evidence (Benson & Clark, 1982; Briggs & Cheek, 1986; Foster & Cone, 1995; Koeske, 1994). For instance, Briggs and Cheek (1986) wrote: “We believe it is best to have an instrument with a known factorial composition before attempting to validate it empirically” (p. 111). Foster and Cone (1995) described validation as comprised of two distinct phases: the *representational* phase and the

elaborative phase. Validity evidence suggesting that a measure adequately represents the construct(s) of interest (e.g., content validity, convergence, and discrimination) were considered major components of the representational phase (Foster & Cone, 1995). Collecting information about the relationships between the measure and variables of practical interest (i.e., test-criterion relationships) represents an important task during the second phase of validation, the *elaborative* phase (Foster & Cone, 1995). These authors acknowledged that evidence collected during the *elaborative* phase could lead to new ideas about the construct(s) of interest, leading to changes to the measure and thus bringing the researcher back to the representational phase of validation. However, the *elaborative* phase of the validation process is more straightforward to the extent that “earlier stages have been marked by both theoretical clarity (i.e., careful definition of the construct) and empirical precision (i.e., careful consideration of psychometric principles and procedures)” (Clark & Watson, 1995, p. 318).

Collecting evidence about the relationships between an instrument and other variables “involves relating scores on a measure to some criterion of practical value” (Foster & Cone, 1995, p. 252) and represents an important way of “elaborating the meaning of scores produced by a measure” (Foster & Cone, 1995, p. 252). The question, in general, when collecting this type of validity evidence is: “How accurately do test scores predict criterion performance?” (AERA/APA/NCME, 1999, p. 14). Of course, the reliability and validity of the criterion are relevant here, just as they are for convergent and discriminant validity (AERA/APA/NCME, 1999). Reliance on a criterion measure that lacks reliability and validity leads the test user to wonder, upon finding a weak association between the measure and the criterion, whether the measure of interest or the criterion variable is to blame (Cronbach & Meehl, 1955). A legitimate

criterion variable, according to Koeske (1994), is one that is measured with essentially no error and that “represents a condition or behavior which is the outcome of practical concern to the tester” (p. 68).

Validity generalization. In education and the behavioral and social sciences, an important issue is whether “evidence of validity based on test-criterion relations can be generalized to a new situation without further study of validity in the new situation” (AERA/APA/NCME, 1999, p. 15). Generalizability, then, pertains to an instrument’s external validity (Cook & Campbell, 1979; Messick, 1993). The failure of the findings of one study to generalize to another study may result from multiple causes, such as: (a) sampling fluctuations and the type of test takers, (b) differences in the range of scores, (c) the type and reliability of criterion variables, (d) differences in the measurement of the constructs, and (e) the time period of the study (AERA/APA/NCME, 1999).

Evidence Based on the Consequences of Testing

The use of tests and other instruments can result in intended and unintended consequences (Messick, 1995). These resulting consequences may be either positive (as in the case of improved services or policies) or negative (such as when bias and unfairness result) (Messick, 1995). Negative consequences may arise from test invalidity, resulting from construct under-representativeness or construct-irrelevant components (AERA/APA/NCME, 1999; Messick, 1993). Seeking evidence that positive consequences have resulted due to the administration of an instrument is a key component of establishing validity based on the consequences of testing (AERA/APA/NCME, 1999). Hayes, Nelson, and Jarrett (1987) referred

to such evidence as the treatment utility of assessment, defined as “the degree to which assessment is shown to contribute to beneficial treatment outcome” (p. 963).

The National Survey of Student Engagement

In the review that follows, several aspects of the National Survey of Student Engagement (Kuh et al., 2001) are discussed and evaluated. First, the conceptual framework and the stated purposes of the NSSE are examined. Then, the items and subscales (benchmarks) comprising the NSSE are discussed, including a critical analysis of the methodology used to construct the benchmarks. Following this discussion is a review of some of the ways institutions of higher education have used NSSE data. Finally, a discussion and critique of the reliability and validity evidence for the five-factor (i.e., benchmark) model of the NSSE is presented.

Conceptual Framework

The National Survey of Student Engagement (Kuh et al., 2001) is a survey administered to first-year and senior-level students at colleges and universities across the nation. The instrument is supposed to measure “the extent to which students are engaged in empirically derived good educational practices and what they gain from the college experience” (Kuh, 2001b, p. 2). Specifically, the NSSE instrument measures *student engagement*, which is defined as: (a) the time and effort students put into their academic studies and other education activities, and (b) the institution’s use of resources and organization of the curriculum and other activities in a way that gets students to participate “in activities that decades of research studies show are linked to student learning” (NSSE, n.d.(a)).

The purpose of the NSSE instrument is three-fold (Kuh, 2009). First, NSSE seeks to provide “high quality, actionable data that can be used to improve undergraduate students’

experiences” (p. 9). According to Kuh (2009), NSSE data serve as proxies for student learning outcomes, and such proxies can lead to institutional and student improvement because of their ability to indicate areas of strength as well as areas of weakness. The second purpose of the NSSE is to learn about and document educational practice in higher education settings (Kuh, 2009). This represents the research-focused goal of NSSE. By including experimental items on the survey and by working with institutions of higher education (Kuh, 2009), NSSE seeks to understand more about student engagement and its effects (c.f., Pascarella and Terenzini, 2005). Finally, NSSE desires to advocate for the public use of data in assessing academic quality (Kuh, 2009). In theory, evidence for a high level of student engagement at a particular institution is evidence that the university is of high quality (NSSE, 2001a). NSSE has created a *Pocket Guide to Choosing a College: Are you Asking the Right Questions* (2008, as cited in Kuh, 2009) for prospective students so that they can “obtain more instructive information about the institutions they are considering” (Kuh, 2009, p. 10). Additionally, researchers at NSSE seek to make their “findings accessible to higher education reporters and the general popular media” (Kuh, 2009, p. 10).

NSSE Items and Benchmarks

According to NSSE, each item on the NSSE instrument serves as “a direct indicator of what students put into their education and an indirect indicator of what they get out of it” (NSSE, 2009b, p.1). Items may be divided into five broad categories:

- student behaviors (e.g., the amount of time spent studying or working off-campus),
- institutional actions and requirements (e.g., the amount of reading and writing required in coursework);

- students' perceptions of the college environment (e.g., the extent to which the institution offers academic support, and the interaction between faculty and students);
- educational and behavioral growth; and
- student demographic information (e.g., gender, race/ethnicity, major) (Kuh, 2009).

Forty-two NSSE items are included in subscales called *benchmarks* (see Table A.1 in the appendix³ for a list of items comprising each of the benchmarks). The five benchmarks comprising the NSSE are: (a) *Level of Academic Challenge (LAC)*, (b) *Active and Collaborative Learning (ACL)*, (c) *Student-Faculty Interactions (SFI)*, (d) *Enriching Educational Experiences (EEE)*, and (e) *Supportive Campus Environments (SCE)*. The five NSSE benchmarks were “created with a blend of theory and empirical analysis” (NSSE, n.d.(d)), and one of the main purposes of them is to facilitate comparisons among institutions (Kuh, 2001a). According to the NSSE technical manual (Kuh et al., 2001):

Educationally effective colleges and universities score above average on all five benchmarks in a manner that is consistent with their mission and students' aspirations and educational goals. Students who are engaged at a reasonable level in all five of these areas gain more than do those who are engaged in only one or two areas (pp. 5-6).

Unfortunately, phrases such as “a reasonable level” of engagement and “gain more” are not elaborated. Two other purposes of the benchmarks are to: (a) provide information that is understandable to a wide audience, such as prospective parents and accreditors, and (b) provide baseline information “against which future performance can be compared” (Kuh, 2001a, p. 14).

³ All tables are displayed in the Appendix.

For the 2009 administration of the NSSE, coefficient alphas for the five benchmarks are provided in Table A.1 (NSSE, 2010a).

The manner in which NSSE benchmarks were created is important to consider. Although NSSE explicitly states that both theory and empirical analysis were used to construct the five benchmarks (Kuh, 2001a; Kuh et al., 2001; NSSE, n.d.(d)), the specific way in which this was accomplished is either not conveyed in the NSSE literature, or is confusing. Kuh (2001a) wrote this about the construction of the benchmarks: “After analyzing the data in different ways we ultimately decided to create five benchmarks based on 40 items” (p. 14). Unfortunately, the NSSE technical guide does not offer much explanation beyond this:

Initially, we conducted principle components analysis with oblique rotation. We also subsequently examined factors produced using an oblique rotation. The results were comparable. Then theory was employed to crystallize the final item groupings (Kuh et al., 2001, p. 30).

Many questions arise from these few sentences: Does this mean that results from a principal components analysis (PCA) were compared to results from an exploratory factor analysis (EFA)? If so, what was the rationale for doing this? How similar were the results? If there were differences, which results did they accept and why? What were the factor loadings for each of the items? What statistical results were produced from the procedures? What is meant by “crystallize the final item groupings” and precisely what theory was used to do this?

That NSSE researchers used PCA *at all* produces many questions about their theory of how the five benchmarks are supposed to be related to the survey items with which they are associated. Floyd and Widaman (1995) described the purpose of PCA succinctly:

...[I]n PCA, ...the components are estimated to represent the variances of the observed variables in as economical a fashion as possible (i.e., in as small a number of dimensions as possible), and no latent variables underlying the observed variables need to be invoked. Instead, the principal components are optimally weighted sums of the observed variables so, in a sense, the observed variables are the causes of the composite variables (p. 287).

When the goal of the researcher is to understand a phenomenon (e.g., a survey) in terms of latent constructs (e.g., subscales), exploratory factor analysis rather than PCA should be used (Brown, 2006; Floyd & Widaman, 1995). Principal components analysis should be reserved for those situations when the researcher wishes to express the variance in a set of observed variables in as few dimensions as possible; in this case, the concept of latent variables is not relevant (Floyd & Widaman, 1995). In an exploratory factor analysis, latent variables are considered to be the underlying causes of the measured variables with which they are associated (Floyd & Widaman, 1995). Conversely, in PCA, observed variables are the causes of the composite variables (Floyd & Widaman, 1995).

The results achieved by using EFA can differ substantially from those found using PCA (Floyd & Widaman, 1995; Widaman, 1993). For example, PCA often yields negative bias in factor covariances, whereas factor covariance estimates produced by EFA are more accurate (Brown, 2006; Widaman, 1993). In addition, factor analysis leads to more accurate estimates of factor loadings, whereas PCA estimates are biased in a positive direction (Widaman, 1993). This is especially likely to occur when only a few observed variables are associated with each factor or when communalities are low (Floyd & Widaman, 1995; Widaman, 1993). Because EFA

results are generally more accurate depictions of population values, and because results from factor analysis are more likely to generalize to those obtained using confirmatory factor analysis (CFA), methodologists generally encourage the use of EFA instead of PCA (Brown, 2006).

Swerdzewski et al. (2007) wrote this about NSSE's use of PCA to construct the benchmarks:

Although the benchmarks were at least partially constructed using PCA, the NSSE literature (NSSE, 2007b) consistently refers to these benchmarks as 'factors' and implies that these benchmarks represent latent traits rather than simple distillations of data as would be accomplished by a PCA approach to data reduction" (p. 5).

To summarize, the use of both PCA and EFA by NSSE researchers creates uncertainty about how the five benchmarks and their corresponding items are hypothesized to relate to one another. By using PCA, NSSE researchers are implicitly stating that the survey items comprising each of the benchmarks are causes of those benchmarks (Floyd & Widaman, 1995). On the other hand, by using common exploratory factor analysis, their implicitly-stated goal was to "discover the latent variables that underlie the scale" (Floyd & Widaman, 1995, p. 286), and in theory, "these latent variables are the underlying causes of the measured variables" (Floyd & Widaman, 1995, pp. 286-287). It simply does not make sense for NSSE items to be both causes *and* effects of the benchmarks.

A related problem of the NSSE benchmarks is that they are generally not homogenous in nature. This problem might be a consequence of the somewhat haphazard way the NSSE benchmarks were initially constructed (Pike, 2006b). In other words, it does not appear that care was taken to ensure that the benchmarks each represented one (and only one) latent variable.

LaNasa, Olson, and Alleman (2007) performed an item-level analysis of their results following analyses based on benchmarks, and “began to view potentially unique sub-components of the benchmark scores as problematic, because without the individual item analysis these differences could go unrevealed” (952). They found it impossible, for instance, to determine whether a low benchmark score for EEE suggested the need for increased diversity, or for more co-curricular activities (LaNasa et al., 2007). Consequently, they created eight subscales (which they called *dimensions*) based on results from a principal factor analysis.

How Institutions Use NSSE

Consistent with the NSSE purpose to “provide actionable data that can be used to improve undergraduate students’ experiences” (Kuh, 2009, p. 9), researchers from NSSE and institutions of higher education have described the use (and potential use) of information gleaned from the NSSE. The uses of NSSE by institutions can be divided into five categories: (a) institutional self-studies, (b) strategic planning, (c) supporting evidence for existing or potential programs and services, (d) assessment and improvement of first-year student courses and programs, and (e) assessment of the engagement of particular groups of students (e.g., minority or transfer students).

Institutional researchers involved in institutional self-studies are encouraged by NSSE to use NSSE survey data as evidence for meeting standards. In fact, NSSE has made it relatively easy for institutional researchers to do this by providing on their website a description of how the NSSE survey items are linked to specific standards for the six regional accrediting associations in the United States (NSSE, n.d.(b)). For example, the item “asked questions in class or contributed to class discussions” may be used as a measure of institutional effectiveness (criteria

3.3.1 of the Southern Association of Colleges and Schools standards). As another example, the item “Worked harder than you thought you could to meet an instructor’s standards or expectations” may be used as a measure of faculty competence (criteria 3.7.1 of the Southern Association of Colleges and Schools standards).

Researchers and campus administrators have also used NSSE data for institutional strategic planning. Pace University, for instance, used NSSE items to assess the extent to which it was meeting the goals expressed in its strategic plan (Kinzie & Pennipede, 2009). Similarly, Grand Valley State University used the NSSE to assess its progress in its “Claiming a Liberal Education” campus change initiative. In addition to using NSSE items to assess institutional strategic planning goals, Oklahoma State University departments and colleges were asked to use results from the NSSE to “identify strengths and potential areas for improvement regarding student engagement” (Banta, Pike, & Hansen, 2009, p. 28).

Additionally, institutions have used NSSE data to gain support for new programs and services for students, and in some cases, to gain support for new staff positions. Elon University, for example, used multi-year NSSE data to identify two areas in which scores had decreased over time (Kuh, 2005). Specifically, fewer students reported having discussed their career plans with a faculty member or advisor, and fewer students reported that they had serious conversations with students from other ethnic backgrounds. In both instances, staff positions were created to specifically address these deficits. At Pace University, NSSE scores were used to make hiring decisions during a hiring freeze (Kinzie & Pennipede, 2009). Because student satisfaction was highly correlated with the quality of academic advising, as well as the quality of relationships with administrative personnel and offices, administration created a new staff position to provide

more centralized services to students through a newly restructured and renamed office of student assistance (Kinzie & Pennipede, 2009).

Because the NSSE is administered to first-year students, survey data can assist researchers in the assessment of programs specific to first-year students. Studies may globally address issues pertinent to the education of first-year students, or they may address institution-specific programs. For example, Zhao and Kuh (2004) examined the social and academic development of first-year students at 364 four-year institutions, and found that participation in learning communities was associated with student satisfaction, outcomes, and engagement. As another example of this type of research, researchers at Southern Illinois University Edwardsville used NSSE results in their evaluation of the quality of different types of first-year seminar courses (Kuh, 2005). Based on the information they obtained, they considered making particular first-year courses required of all students. Similarly, Austin Peay State University used data to develop and refine first-year seminars and orientation workshops (Kinzie & Pennipede, 2009). At other institutions, faculty members have added NSSE items to their end-of-year course evaluations to investigate ways to improve student engagement through their teaching (Kinzie & Pennipede, 2009). Additionally, stakeholders at Indiana University—Purdue University Indianapolis matched specific items from the NSSE to their program goals, and plan to use NSSE results to track their improvement toward these goals (Banta et al., 2009, p.24).

Finally, NSSE results have been used to assess the engagement of specific types of students. Past research suggests that in general, certain groups of students (e.g., women, full-time students, students who live on campus, non-transfer students, students who participate in learning communities, and international students) have higher scores on the NSSE (NSSE, 2000). Some

institutions have used NSSE data to assess the engagement of student athletes or traditionally underrepresented students (Ahren, Ryan, & Massa-McKinley, 2008).

To summarize, NSSE data are being used in various ways by institutions across the nation. As the NSSE literature has suggested, many institutions of higher education are basing their decisions and policies on NSSE results. Thus, it seems especially important to evaluate the validity of the instrument. This discussion now turns to the literature that has focused on the reliability and validity of the NSSE.

Reliability of NSSE Benchmarks

Many studies have attempted to shed light on the reliability of the NSSE. Based on Kane's (1992) argument-based approach to validity, "the evidence needed to support assumptions about invariance is collected in reliability studies...which indicate how consistent scores are across different samples of observations (e.g., across different samples of items, occasions)" (Feldt & Brennan, 1989, p. 529). Reliability evidence supporting these invariance assumptions may be organized into three broad categories: (a) alternate-forms reliability, (b) test-retest reliability, and (c) internal reliability (AERA/APA/NCME, 1999). Recall that when a subscale or scale is thought to represent a one-dimensional construct, evidence for internal reliability (i.e., consistency across items) also provides evidence for validity (specifically, it provides evidence for internal structure). Thus, studies that report interitem correlations, coefficient alphas, results from factor analyses, or results from confirmatory factor analyses are discussed in the validity portion of the review. It is important to point out that all these sources of error are cumulative, and that the sources of reliability that the researcher chooses to explore

should depend on the most likely or relevant sources of error (Anastasi & Urbina, 1997; Cronbach, Gleser, Nanda, & Rajaratnum, 1972).

The NSSE Institute conducted an analysis to determine the test-retest reliability for each of the benchmarks (NSSE, 2010b). To do this, they used data collected from (a) students who responded to both the paper and the online version of the survey, and (b) students who responded to the paper version of the survey on two separate occasions. For the 2009 administration of the survey, the test-retest reliabilities of the five benchmarks ranged from .62 to .68 for first-year students, and from .66 to .76 for seniors (NSSE, 2010b). The amount of time elapsed from one administration to the other was not reported.

Validity of NSSE Benchmarks

As stated earlier, validity is a broad concept (AERA/APA/NCME, 1999; Anastasi & Urbina, 1997; Messick, 1993). This review of the NSSE validation literature is organized using the broader categories described in the Standards (AERA/APA/NCME, 1999) (e.g., evidence based on test content). In many instances, researchers sought to obtain more than one “type” of validity evidence (e.g., Gordon, Ludlum, & Hoey, 2008). Such studies are organized by the *primary* type of evidence sought. On the other hand, some types of validity evidence were not explored in any of the literature; in particular, no studies sought out evidence for validity generalizability or evidence based on the consequences of testing.

It is important to note that this review is limited to those studies using the NSSE benchmarks exactly as created.⁴ There are many studies linking individual items or various groups of items to criteria such as student achievement or satisfaction (e.g., Carini, Kuh, &

⁴ Some studies reviewed did, however, drop a few items for various reasons. For instance, LaNasa, Cabrera, & Trangsrud (2009) deleted five items from their analysis because responses were skewed, and/or items were measured on a nominal scale of measurement.

Klein, 2006; Pascarella et al., 2008). Such studies are not described here because they provide weaker evidence for the validity of the benchmarks as they were created by NSSE.

Test content. When developing a measure, it is important to first “develop a precise and detailed conception of the target construct and its theoretical context” (Clark & Watson, 1995, p. 310). As stated earlier the NSSE items were “specifically designed to assess the extent to which students are engaged in empirically derived good educational practices and what they gain from their college experience” (Kuh, 2001b, p. 2). A reader questioning, “What empirically derived good educational practices?” is not alone. Porter (2009) ascribed this lack of clarity to the failure of NSSE researchers to state the content domain clearly: “...the domain is so widely defined that almost any student survey question could be included under the areas ‘engagement,’ ‘student outcomes,’ and ‘institutional quality’” (p. 11). Another point made by Porter (2009) was that survey items “are included because they are “‘arguably’...related to student outcomes as shown by research” (p. 11), yet there is no explanation offered by NSSE as to how items are linked to specific constructs of interest (or even what those specific constructs of interest are) (Porter, 2009). Items are included on the survey because they represent “behaviors that are highly correlated with many desirable learning and personal development outcomes of college” (Kuh, 2001b, p. 2). No description was provided by NSSE, however, for “desirable learning and personal development outcomes.” Kuh (2001b) asserted that NSSE researchers “devoted considerable time...making certain the items on the survey were clearly worded, well-defined, and had high face and content validity” (p. 5). Unfortunately, these efforts were not well-documented as suggested in the validation literature (e.g., Benson & Clark, 1982), and one is left to wonder how, exactly, survey items were created or selected for inclusion in the final

instrument. A related question is, What constructs do the benchmarks represent? Explanations of NSSE benchmarks generally reference the specific items included in the benchmarks (e.g., Pascarella, Seifert, & Blaich, 2010), but fail to provide a clear description of, for instance, the *construct* referred to as “enriching education experiences.”

Recall that the word “content” in the phrase “validity based on test content” refers not only to the items comprising an instrument, but also to the formatting of the measure (AERA/APA/NCME, 1999). Most of the NSSE benchmark items are answered on a four-point scale (e.g., frequency questions are answered on a 4-point scale ranging from “very often” to “never”). According to Benson and Clark (1983), an important part of developing an instrument is choosing item format. What is not clearly stated in any of the NSSE literature is *why* questions are answered using this type of response set. That only four responses are possible for respondents is unfortunate because of the problems ordinal data may present for statistical analyses, *especially* when there are less than seven or so categories (e.g., violation of normality assumption; Tabachnick & Fidell, 2007).

Response processes. “Theoretical and empirical analyses of the response processes of test takers can provide evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by examinees” (AERA/APA/NCME, 1999, p. 12). In their study, Ouimet et al. (2001) used focus groups of students to gain evidence for the validity of the NSSE. Their participants included 221 students in 35 separate focus groups at eight colleges and universities. In addition to overall clarity, the authors sought information about (a) whether students interpreted NSSE items in a manner consistent with what was intended, and (b) exactly what students meant when they answered items pertaining to the

frequency of behaviors; for instance, what is the distinction between “very often” and “often”? These focus groups allowed Ouimet et al. (2001) to discover issues with the survey’s appearance (e.g., it looked like a test), item clarity, and response options. In many cases, students’ input about item clarity was backed with statistical evidence that there might be a problem. For instance, in many cases, items identified as problematic in focus groups were not normally distributed or had low item-scale correlations. Feedback from the focus group resulted in wording revisions for 12 NSSE items. By changing these items, the authors’ goal was to “increase the clarity and measurement precision of the items on [the NSSE], thereby increasing the instrument’s validity and reliability” (p. 16). Unfortunately, follow-up analyses were not conducted to see whether changes resulted in any improvements in the survey instrument.

Regarding their second research question—concerning the way students interpreted frequency phrases such as “very often” and “often”—Ouimet et al. (2001) found that students defined these terms differently depending on the item being asked. The authors asserted that students showed agreement for most of the individual items, however. For example, they agreed that for the item, “talked about career plans with a faculty member or advisor,” *occasionally* meant “a couple times a year.” The authors provided a table indicating, for each frequency item, the definitions of “very often,” “often,” “occasionally,” and “never.”

Internal structure. Most studies of the NSSE include internal consistency measures such as coefficient alpha. On the other hand, only two studies have examined the factorial validity of the NSSE as represented by a five-factor (benchmark) model. In the most recent study, LaNasa et al. (2009) sought to examine the factorial validity of the NSSE benchmarks by conducting a confirmatory factor analysis. The sample they analyzed consisted of 375 first-year students

(about 60% female) in 2004. The institution was described as a public, doctoral, research-intensive university in the Midwest region. Before performing the confirmatory factor analysis, internal reliability was assessed; estimates ranged from $\alpha = .59$ (*EEE*) to $.79$ (*SCE*).

The overall fit of the model was described as reasonably acceptable ($\chi^2/df = 2.70$, $TLI = .92$, $CFI = .93$, $RMSEA = .07$). The values of these fit statistics fall slightly out of the range of recommended criteria, however (Hu & Bentler, 1999). In addition, some items failed to load onto the expected benchmarks, and the correlation between two benchmarks (*ACL* and *SFI*) was high enough ($r = .89$) to warrant concern about discriminant validity. In addition, for the *ACL* benchmark, none of the seven indicator variables had loadings of $.70$ or larger. Only 11 of the 30 remaining items had high factor loadings (greater than $.70$). Altogether, these results paint a picture of the factorial validity of the NSSE benchmarks that merits concern. The authors of this research concluded that “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” (LaNasa et al., 2009).

To follow their own advice, LaNasa et al. (2009) performed a principal components analysis that yielded a nine-factor solution explaining 61% of the variance. They named their constructs: (a) Learning Strategies, (b) Academic Interaction, (c) Institutional Emphasis, (d) Co-curricular Activity, (e) Diverse Interactions, (f) Effort, (g) Overall Relationships, (h) Workload, and (i) Working Collaboratively In-class. Unfortunately, the authors did not state why they used PCA rather than exploratory factor analysis. After performing a confirmatory factor analysis with the resulting factors, they modified their model to reflect only eight factors (i.e., they deleted the Working Collaboratively In-class factor along with its two corresponding items).

Similar to the research by LaNasa et al. (2009) just described, Swerdzewski et al. (1997) at James Madison University assessed the factorial validity of the NSSE benchmarks. They used confirmatory factor analysis to answer their first research question, “Will an internal study per Benson’s (1998) second phase of a strong program of construct validity yield acceptable evidence for the factor structure of the five cross-institutional benchmarks specified for the NSSE?” (p. 9). Using 2005 data from 495 full-time first-year students (73% were female, and 83% were Caucasian), Swerdzewski et al.’s (1997) confirmatory factor analysis results indicated poor model fit ($\chi^2/df = 5.31$, CFI = .53, RMSEA = .09). Unfortunately, information about individual item loadings was not provided, but the overall message is consistent with that of LaNasa et al. (2009). These authors concluded that “a comparison of benchmark scores from this sample to scores from a sample at another university should not be made, as the benchmark scores from this sample were not empirically supported” (p. 16). A further word of advice from these authors is worth mentioning here because it is well-stated and clear: “Without adherence to a strong program of construct validity, inferences made from instruments that are even as well-known as the NSSE must be made with extreme vigilance” (LaNasa et al., 2009, p. 18).

Relations to other variables. The literature review now turns to a discussion of research that has explored the relationships between the NSSE (as depicted by the five benchmarks) and other variables. For reasons described earlier, only those studies employing criterion variables that were external to the NSSE itself are reviewed. Results of such studies are largely dependent upon the psychometric properties of the criterion variables, as well as the psychometric properties of the NSSE (Koeske, 1994).

Convergent and discriminant validity. Little evidence has been amassed to show that the NSSE is related to other measures as expected; in other words, there is not enough evidence to suggest that student engagement, as depicted by the five benchmarks, is related to similar or different constructs in predictable ways. Kuh (2001b) implied that because most items on the NSSE were pulled from “other long-running, well-regarded” (p. 4) research programs and questionnaires [e.g., UCLA’s Cooperative Institutional Research Program (Astin, 1993; Sax, Astin, Korn, & Mahoney, 1997), and Indiana University’s College Student Experiences Questionnaire Research Program (Kuh, Vesper, Connolly, & Pace, 1997; Pace, 1984, 1990)], that the NSSE is a valid instrument. To make this claim about the NSSE’s validity, one would need to know the extent to which other measures are correlated with the NSSE. Cronbach and Meehl (1955) wrote:

If a new test is demonstrated to predict the scores on an older, well-established test, then an evaluation of the predictive power of the older test may be used for the new one. But accurate inferences are possible only if the two tests correlate so highly that there is negligible reliable variance in either test, independent of the other (p. 285).

Clearly, more validation studies are needed that focus on the NSSE. The argument that the NSSE has convergent or discriminant validity based on the fact that its content is similar to other instruments is weak at best.

Test-criterion relationships. Test-criterion validity evidence is desirable when the information gleaned from an instrument or other assessment tool will be used to make predictions (e.g., Benson and Clark, 1983). Nowhere in the NSSE literature is reference made to the NSSE as a prediction-making tool; however, much research has been carried out to see how

NSSE results are related to criteria such as GRE scores and college GPA (e.g., Carini et al., 2006). This research is relevant to the validity of the NSSE because student engagement is theoretically tied to student outcomes such as learning and personal development (e.g., Kuh, 2009). In fact, "...one of the major assumptions of the NSSE is that in measuring the extent to which students engage in such practices, one is indirectly measuring student cognitive and personal development during college" (Pascarella et al., 2010, p. 18). Thus, to the extent that the NSSE is associated with such outcomes, validity evidence is accumulated for the construct of student engagement.

Carini et al. (2006) conducted a study to see if student engagement, as measured by the NSSE, was related to three outcomes after controlling for SAT scores: (a) scores on an academic performance test (the RAND), (b) GRE essay scores; and (c) cumulative college GPA. In this study, students from all class levels (e.g., sophomores) were included. Carini et al. (2006) computed the bivariate and partial correlations between the five NSSE benchmarks and these three outcomes. Overall, findings were less than optimal in terms of supporting the validity of the benchmark model of the NSSE. Specifically, three of the five correlations between the five NSSE benchmarks and the RAND test were statistically significant, two of the five correlations between the benchmarks and GRE scores were significant, and all of the five correlations between NSSE benchmarks and cumulative GPA were significant. One might expect *all* of these correlations (that is, 100% instead of 67%) to be statistically significant if (a) the benchmarks are representative of student engagement, (b) the criterion measures used were valid, and (c) student engagement is truly linked to student outcomes as proposed in the literature (e.g., Pascarella & Terenzini, 2005). After controlling for SAT score and other student demographic information

(e.g., class, gender, enrollment status), the number of statistically significant correlations between NSSE benchmarks and student outcomes declined: two benchmarks were correlated with the RAND test, two benchmarks were correlated with GRE scores, and four benchmarks were correlated with cumulative GPA. This means that, after controlling for other things that might influence the three outcomes of interest in this study, just over 50% of the correlations between the five NSSE benchmarks and the three investigated student outcomes reached statistical significance. Effect sizes for these findings were very small, small enough for the authors to conclude that “a large portion—and in some cases a majority—of the variance in key outcomes remains to be explained by yet undiscovered factors” (p. 23). To summarize Carini et al.’s (2006) study concisely, NSSE benchmarks were generally not valid predictors of the student outcomes explored, perhaps with the exception of students’ GPA.

Using multiple regression, Gordon et al. (2008) sought to obtain evidence for the test-criterion validity of the NSSE benchmarks at their institution. To do this, they determined whether the five NSSE benchmarks predicted four student outcomes after controlling for student characteristics such as gender, ethnicity, major, and parental education. The four student outcomes they examined were: (a) cumulative college GPA (first year students and seniors examined separately), (b) first-year retention (first-year students only), (c) job placement following graduation (seniors only), and (d) the intent to enroll in graduate school (seniors only). Prior to performing the five regression analyses, the authors examined the internal consistency of the benchmarks using coefficient alpha and found these values to range from .57 to .70 for first-year students and from .53 to .70 for seniors. The authors noted that these ranges represent unacceptable to minimally unacceptable internal consistency. The results from the first multiple

regression indicated that for first-year students, three benchmarks were significantly associated with cumulative GPA when controlling for the other variables in the model: *LAC*, *ACL*, and *EEE*. Unexpectedly, the *EEE* benchmark was *negatively* associated with cumulative GPA ($\beta = -.11$). For seniors, only the *SCE* benchmark was a significant predictor of cumulative GPA when demographic variables and the other benchmarks were controlled. In the third multiple regression, first-year retention of first-year students was regressed on the control demographic variables and the five NSSE benchmarks. Of the five benchmarks, only the *SCE* benchmark was statistically significant. In the fourth multiple regression, seniors' employment placement at the time of graduation was regressed on the demographic variables and the five NSSE benchmarks, none of which emerged as statistically significant. In the fifth and final multiple regression, SFI was a significant predictor of seniors' intention to enroll in graduate school when the other variables were controlled. Similar to the results of Carini et al. (2006), this study did not provide strong evidence for the validity of using NSSE benchmarks to predict several student outcomes.

In another study, Pascarella et al. (2010) sought to examine the predictive validity of the NSSE by estimating the correlations between scores on each of the NSSE benchmarks and five measures of liberal arts outcomes (Pascarella et al., 2008; 2010). Specifically, the outcomes they were interested in were: (a) effective reasoning and problem-solving, (b) moral character, (c) inclination to inquire and lifelong learning, (d) intercultural effectiveness, and (e) personal well-being. The researchers were interested in two levels of analysis: the institution level and the individual level. In estimating correlations at the institutional level, they statistically controlled for pre-existing student characteristics with the mean score of a pre-test (which was the same instrument as that used to measure the liberal arts outcomes). By using a pre-test, the authors

sought to minimize the effects of student characteristics that might be associated with each institution's recruitment and admission policies (Pascarella et al., 2010). Of the 35 correlations they computed (5 benchmarks x 7 outcome measures), only 10 (29%) were statistically significant.⁵ Although the authors did not make explicit the statistical alpha level used in their 2010 article, they wrote in their 2008 unpublished manuscript that due to the small sample of institutions ($n = 19$), they used an alpha level of .10. Based on this alpha level, nearly four items would be statistically significant due to chance alone. Given that only 10 of the 35 correlations were significant, and the fact that Type I error risk was increased due to the multiple statistical tests,⁶ the authors' conclusion was surprising: "...we concluded that institution-level NSSE benchmark scores had a significant overall positive association with the seven liberal arts outcomes at the end of the first year of college..." (Pascarella et al., 2010, p. 20). The authors correctly noted that their power was low due to having a small sample, but to conclude that "the dimensions of the undergraduate experience measured by NSSE benchmarks are correlated with important educational outcomes [and] arguably constitute a more valid conception of quality in undergraduate education than *U.S. New's*" (Pascarella et al., 2010, p. 21) seems quite unsubstantiated.

Pascarella et al. (2008) also performed analyses at the individual level. Their sample consisted of 3,081 first-year, full-time students at 19 institutions. The authors computed correlations between four of the NSSE benchmarks (EEE was dropped from the study because of low internal consistency) and 15 measures representing the five dimensions of student outcomes

⁵ In their 2008 study, there was a total of 15 criterion variables. Of the 75 partial correlations they computed, only 16 (21%) were statistically significant at the .10 alpha level.

⁶ Because of the multiple significance tests, the probability of making at least one Type I error (i.e., "false positive") is $1 - (1 - .10)^{35} = .97$ (Abdi, 2007).

previously listed (e.g., moral character). They computed two sets of correlations: one adjusting for scores on the other benchmark scales (which they termed “upper bounds estimates”), and one adjusting for other benchmark scores in addition to a pretest of the outcome measures, precollege academic preparation, race, sex, parental education, precollege academic motivation, high school involvement, place of residence, work responsibilities during college, liberal arts emphasis of first-year college coursework, and institutional type (they referred to this set of correlations as “lower bounds estimates”). Interestingly, the authors cited evidence (Bowman, 2010) that pulling criterion variables from the NSSE instrument itself would be problematic because all variables are self-reported by students. However, they overlooked the fact that most of the criterion variables they relied upon were also based on self-reported perceptions and attitudes, and the fact that those criterion variables may have resulted in inflated correlations with benchmarks because of their shared method variance (Koeske, 1994; Messick, 1993).

Of the 60 correlations that controlled only for other NSSE benchmark scores, 34 (57%) were statistically significant at the .05 significance level (significant *rs* ranged from .09 to .28). For their more conservative “lower bounds estimates” that included multiple control variables, 29 of the 60 correlations (48%) were statistically significant (significant *rs* ranged from .04 to .15). Even though only about half of the correlations for each set of analyses were statistically significant, the authors concluded that their results “confirm the primary assumption of NSSE that in measuring the student exposure to and engagement in empirically-vetted good practices, one is essentially measuring experiences that yield desired cognitive and personal development during college” (Pascarella et al., 2008, p. 3).

NSSE Scalelets

The benchmarks used by NSSE to describe student engagement may not be useful for researchers and administrators seeking to use NSSE data to understand the state of affairs at their institutions, or to improve services and programs. For this reason, Pike (2006b) suggested a different representation of the NSSE items using what he called *scalelets*. According to Pike (2006b), scalelets are “clusters of highly related survey questions representing the experiences of a group of students” (p.178) that are more likely to enable the institution to use NSSE data. Specifically, the 12 scalelets purportedly make it easier for institutions to determine areas of strength and weakness, and to take action accordingly. Pike (2006b) further described a scalelet as consisting of “a limited number of survey questions that provide a measure of a specific aspect of the educational experience of a group of students” (p.181). The contents of the NSSE scalelets generally parallel the content of the NSSE benchmarks; that is, scalelets may simply be thought of as smaller subdivisions of the original five NSSE benchmarks. More NSSE items are included, however, in Pike’s scalelet model ($n = 49$), than in NSSE’s benchmark model ($n = 42$). Pike’s (2006b) scalelets and their corresponding survey items are displayed in Table A.2.

Only two studies exist that explore the psychometric properties of the NSSE scalelets (Pike, 2006a; 2006b). However, some of the validity evidence pertaining to the NSSE benchmarks is also applicable to the scalelets. For instance, the procedures used to create items (e.g., Kuh, 2001b; Pascarella et al., 2010) and evaluate the processes by which students answer NSSE items (e.g., Ouimet et al., 2001) should also generally apply to NSSE scalelets because scalelets are simply smaller sets of items that appear on the NSSE instrument itself. Thus, what follows is information pertaining *specifically* to the scalelets that has not been presented earlier.

Pike (2006b) used generalizability theory⁷ (Cronbach et al., 1972) to explore the dependability of 12 newly-created scalelets “based on face and content validity” (Pike, 2006a, p.553). The first research question asked whether group means would “generalize to the universes of all possible items and all possible students” (Pike, 2006b, p. 187). Based on a random sample of 50 seniors, all from different institutions, Pike (2006b) concluded that four (perhaps five) of the scalelets (Course Interaction, Out-of-Class Interaction, Varied Experiences, Support for Student Success, and perhaps Higher Order Thinking Skills) could produce dependable (reliable) group means “that could be generalized to all possible items and students based on a sample of 50 students” (p. 187). On the other hand, four scalelets (Writing, Collaborative Learning, Information Technology, and Interpersonal Environment) would not provide reliable generalizations even if the sample size was increased to 200. Although only 42% of the scalelets were found to be dependable with a sample size of 50, and 67% were found to be dependable with a sample size of 200, Pike (2006b) concluded that “NSSE scalelets produce highly dependable group means based on relatively small samples of students when the assessment researcher is interested in making judgments about the mean (or sum) of items comprising the scalelet” (p. 190). Pike did, however, adequately describe the unreliability of some of the scalelets as the result of the way items were selected for the NSSE instrument:

The questions...were not selected to represent random samples from larger domains.

They were initially selected for the survey because of their face validity and because they were related to good educational practices...Given the absence of an intentional sampling

⁷ The focus of generalizability theory is on the errors from multiple sources as separate entities, in contrast to classical test theory that considers measurement error as a single entity arising from multiple sources (Feldt & Brennan, 1989). By partitioning error in this manner, the magnitude of the primary sources of measurement error can be assessed.

frame, it is not surprising that the items included in the scalelets represented rather poor samples of particular constructs (p. 191).

Pike's (2006b) second research question asked whether group means would generalize when the goal was "to draw conclusions about the items comprising a scalelet, rather than a higher-order construct" (p. 187). In other words, when the objective is to generalize to the universe of students (but not items), what sample size is necessary? Pike found that all 12 scalelets were dependable in this respect with a sample size of 50.

Pike (2006a) continued his research on his newly-created NSSE scalelets by evaluating whether they were related to self-reported student gains.⁸ Items were drawn from two scales: the Gains in General Education scale, and the Gains in Practical Competence scale (Kuh, Gonyea, & Palmer, 2001). No psychometric properties are available for either of these scales. The Gains in General Education scale included items about gains in writing, speaking, analytical skills, and general education. The Gains in Practical Competence scale required students to report their gains in computer and information technology, quantitative skills, and work-related skills and knowledge.

To see if, and to what extent, the NSSE scalelets were related to student-reported gains, two multiple regression models were tested, with institutions serving as the unit of analysis.⁹ First, the Gains in General Education scale was regressed on institutional characteristics (i.e., institutional type, Carnegie classification, percent of female students, percent of minority students, percent of on-campus students, and percent of full-time students) and the 12 scalelet

⁸ Although Pike (2006a) referred to his study as one investigating convergent validity, the focus of his study is more appropriately described as one examining test-criterion validity.

⁹ Pike (2006a) also ran two more multiple regressions, using benchmarks instead of scalelets. He found that benchmarks accounted for 31% of the variance in the Gains in General Education scale, and 22% of the variance in the Gains in Practical Competence scale, after controlling for institutional characteristics.

scores. Then, in a separate regression, the Gains in Practical Competence scale was regressed on institutional characteristics and the 12 scalelets. Hierarchical regression was used to see if the scalelets explained variance in the self-reported gains beyond that which could be explained by institutional characteristics alone. Together, institutional characteristics and NSSE scalelets explained 81% of the variance and 54% of the variance in The Gains in General Education scale, and the Gains in Practical Competence scale, respectively. When institutional characteristics were already accounted for, the 12 scalelets accounted for an additional 34% and 36% of the variance in the Gains in General Education scale and the Gains in Practical Competence scale, respectively.

Two serious limitations in this study were discussed by Pike (2006a). First, gains in student learning were measured with self-reported data. Like much of the NSSE literature, this is presented as unproblematic because “self-report data have been studied extensively and shown to yield valid assessment information (see Kuh; 2001)” (p. 557). However, there is recent evidence to suggest that the correlations between student-reported gains and *actual* gains are quite small (Bowman, 2010). A second limitation is one already discussed, one of shared method variance (Koeske, 1994; Messick, 1993). All measures of interest in the study (i.e., gains and NSSE scalelets) were collected using self-report. This means that the ability of the scalelets to explain variance in the self-reported gains may have occurred simply because the measures were all collected using self-report survey instruments. A third problem Pike (2006a) did not address is the lack of data to support the reliability and validity of the two gain measures. When relating tests to outcome measures, it is imperative that the outcome measures be psychometrically valid; otherwise, results are inconclusive (Cronbach & Meehl, 1955). To summarize using Pike’s

(2006b) words: “Additional research, focusing on the construct validity of scalelets, is required before concluding that they should be used in assessment research” (Pike, 2006b, p. 190). The current research represents such a study by evaluating the factorial validity of Pike’s (2006b) scalelet model.

Summary of Literature Review

The purpose of the preceding literature review was to discuss existing research in three areas: (a) instrument validation, (b) the NSSE benchmarks, and (c) Pike’s (2006b) scalelet model of the NSSE. The summary suggests that there are many ways to examine the validity of an instrument, and that the various ways altogether provide evidence in support of or against an instrument’s use. To date, only two studies have examined the factorial validity of NSSE benchmarks and none have looked at the factorial validity of Pike’s (2006b) NSSE scalelets. Such an investigation is important, however, to determine whether survey items are representative of the factors with which they are meant to be associated. The focus of the paper now turns to the current study. In the next chapter, the method used to collect and analyze the data for this study is described in detail.

CHAPTER 3: METHOD

The statistical analyses using data from the 2009 NSSE administration addressed three research questions. First, does a five-factor model of student engagement (i.e., the benchmarks) adequately represent the data? A second related question pertains to whether Pike's (2006b) scalelet model of student engagement adequately represents the data. Because neither of the two models were satisfactory, a third research question was addressed: Is there a model that better fits the data for The University of Tennessee, Knoxville?

Instrument

All variables of interest in this study were drawn from the 2009 National Survey of Student Engagement. The instrument contains 85 questions, plus 24 demographic items. As previously discussed, items pertain to student behaviors and perceptions, as well as institutional actions and behaviors (Kuh, 2009). Two compositions of NSSE data were of interest in the current study: the five-factor benchmark model proposed by NSSE, and Pike's (2006b) scalelet model.¹⁰

Procedure

Most administration issues associated with the National Survey of Student Engagement were handled by the NSSE Institute. The sample of first-year students and seniors ($n = 5,000$) invited to participate in the online survey represented approximately 46% of students in those two classes ($n = 10,845$) at The University of Tennessee, Knoxville in 2009. An email inviting students to complete the NSSE online was sent to students in early February, with a follow-up email sent to students one week later. Three reminders were emailed to students in late February,

¹⁰ The original scalelet model (Pike, 2006b) is comprised of 12 subscales. For this study, the model necessarily needed to be reduced to nine subscales because once items were deleted (because of their nominal scale of measurement or non-normality), three scalelets were associated with less than three indicator variables.

early March, and late March. Students who chose to participate in the study were entered into a drawing for one of five gift cards (three \$100 denominations and two \$50 denominations) for the university's dining facilities.

The Office of Institutional Research and Assessment (OIRA) was responsible for maintaining NSSE data. Permission was granted by the director of OIRA to use NSSE data for this study, and IRB approval was also granted by the Institutional Review Board at The University of Tennessee, Knoxville.

Participants

The response rate for the 2009 administration was 38.5%. The participants of this study were first-year ($n = 981$) and senior ($n = 944$) students at The University of Tennessee, Knoxville (UT) who completed the NSSE online in the spring of 2009. The sample of females was higher ($n = 1,157$, or 60%) than that of males ($n = 768$, or 40%), even though females represent about half (49%) of the population of students at UT. Consistent with the population of UT students, the majority of respondents was Caucasian (83%). Six percent of students were Black or African, 3% were Asian, Asian American, or Pacific Islander, and 2% were multiracial. Fewer than 1% of UT student claimed other ethnicities (e.g., .5% were American Indian or other Native American). About 4% of students answered "I prefer not to respond" to the ethnicity question.

Confirmatory Factor Analysis

Structural equation modeling is a statistical method that allows one to analyze the relationships among multiple factors (constructs) (Byrne, 2010). Structural equation models are comprised of two sub-models: (a) measurement models, and (b) structural models¹¹. The

¹¹ Structural models, which are not used in the current study, concern the relationships among latent factors (Byrne, 2010).

measurement model represents “hypotheses about the relations between a set of observed variables, such as ratings or questionnaire items, and the unobserved variables or constructs they were designed to measure” (Hoyle & Smith, 1994, p. 429). It is the measurement model (and not the structural model) that is of importance in the current study because it corresponds to the confirmatory factor analysis model (Brown, 2006; Byrne, 2010). For an example of a measurement model, refer to Figure 3.1.

Confirmatory factor analysis (CFA) may be used to acquire validity evidence for a measure (Brown, 2006; Byrne, 2010; Harrington, 2009). Specifically, it can be used to evaluate the factorial validity of an instrument such as a survey (Hoyle & Smith, 1994). It is statistically similar to exploratory factor analysis (EFA) because it is based on the common factor model. Like EFA, CFA partitions the variance of observed variables into two groups: shared variance and unique variance (Brown, 2006). Shared variance refers to the variability of an observed variable that is shared with the latent variable, whereas unique variance refers to the variance of an observed variable that is *not* shared with the latent variable with which it is associated (Brown, 2006). Unlike EFA, however, CFA requires the researcher to posit a model a priori based on some combination of theory and previous research (Briggs & Cheek, 1986; Byrne, 2010). This hypothesized model is then evaluated for plausibility by comparing it to sample data that have been collected (Briggs & Cheek, 1986). Confirmatory factor analysis is particularly useful “when it is used to assess both whether a proposed factor structure adequately fits the data and whether the structure fits as well as and as parsimoniously as other models” (Floyd & Widaman, 1995, p. 293). For example, CFA may provide information about the number of factors (or subscales) that comprise an instrument, and whether these factors,

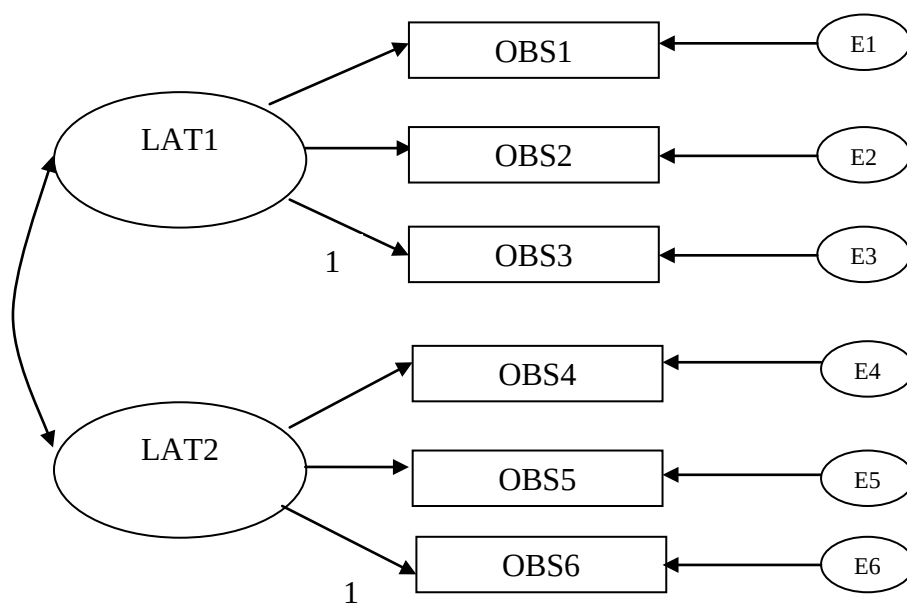


Figure 3.1. Measurement model. LAT1 and LAT2 represent two correlated latent variables. There are a total of 6 observed variables (e.g., OBS4). Errors are not allowed to covary. The model has been scaled by setting OBS3 and OBS6 to equal 1.

combined, are reflective of a higher-order factor or scale (Brown, 2006). The process of using CFA to examine the plausibility of a model may be broken into three steps: (a) model specification, (b) model estimation, and (c) evaluation of model fit (Harrington, 2009). Each of the three main steps of CFA is discussed below, as it pertains to the validation of an instrument such as NSSE.

Model specification. As mentioned above, confirmatory factor analysis requires the researcher to hypothesize a model for the data based on theory or prior research before the analysis has begun (Briggs & Cheek, 1986; Byrne, 2010). This involves specifying the observed and latent variables in the model. The observed variables (such as *OBS1* in Figure 3.1) are survey items, observations, or some other variable that is measured by the researcher. Latent variables, on the other hand, represent the underlying constructs of interest to the researcher. In Figure 3.1, there are two latent variables: *LAT1* and *LAT2*. The researcher may believe that items on a survey cluster into subscales, for instance. These subscales, although of interest to the researcher, are not directly measured but their values are inferred from the observed variables. In addition to specifying the observed and latent variables in a model, the researcher must also specify whether the latent variables are presumed to correlate with one another. The curved line between the two latent variables in Figure 3.1 signifies such a correlation.

Sometimes, the researcher may also specify that correlated error variances (of the observed variables) are expected as the result of, for instance, the common methods used to collect data or the common wording of items (Harrington, 2009). Unlike EFA, confirmatory factor analysis allows for the modeling of such correlated error variances (Brown, 2006). Recall that CFA partitions item variances into two groups: that which is shared with the latent variable,

and that which is not shared (unique). Correlated error variances represent correlations between items that exist because of some reason other than the shared influence of the latent variable (Brown, 2006; Harrington, 2009). For instance, items may be correlated simply because they are worded similarly or deal with the same content (Brown, 2006; Floyd & Widaman, 1995; Harrington, 2009). In CFA, when correlated errors are not posited, “[t]he underlying assumption...is that the observed relationship between any two indicators loading on the same factor...is due entirely to the shared influence of the latent dimension...” (Brown, 2006, p. 46). Usually, specification of correlated error variances is post-hoc in nature; that is, the researcher initially proposes a model with no error covariances and then, based on the results of the analysis, adds error covariances when it appears that doing so would result in a better-fitting model (Hoyle & Smith, 1994). Importantly, one must be sure when adding error covariances that doing so makes sense theoretically; otherwise, the error covariances are probably due to chance and are not likely to replicate in future studies (Brown, 2006). In Figure 3.1, errors are specified to not be correlated because there is no curved line between any of the error terms (e.g., between *E1* and *E2*).

Identification. While specifying a model to be analyzed using confirmatory factor analysis, one needs to be sure that a unique solution is possible by ensuring that two conditions have been met: (a) the model should have more known parameters than unknown ones (i.e., degrees of freedom must be greater than 0), and (b) latent variables need to be scaled (Brown, 2006; Kline, 2005; Harrington, 2009). The second condition, that latent variables need to be scaled, means that a unit of measurement must be established for latent variables which do not already have measurement units because of their latent nature (Brown, 2006; Harrington, 2009).

To scale latent variables, there are two possibilities: (a) make the unit of measurement the same as one of the indicator variables with which it is associated (called a *reference indicator*), or (b) set the variance of the latent variable to equal 1 (Harrington, 2009). The procedure used generally does not affect results of the analysis, although differences in the two procedures are possible (Brown, 2006; Harrington, 2009). The measurement model depicted in Figure 3.1 was scaled using the reference indicator method.

Model estimation. The basic task of confirmatory factor analysis using structural equation modeling is to see if the model, as specified by the researcher, fits the data at hand. More specifically, the researcher wants to determine if the variance-covariance matrix (Σ) of a hypothesized model is similar to the variance-covariance matrix (S) of the sample data. In confirmatory factor analysis, there are multiple estimation methods to determine whether the model fits the sample data. The most common of these methods is the maximum likelihood (ML) method. Brown (2006) described the goal of ML estimation as finding “the model parameter estimates that maximize the probability of observing the available data if the data were collected from the same population again...ML aims to...maximize the likelihood of the parameters given the data” (p. 73). Like most other inferential statistics, the ML method requires certain conditions to be met. Specifically, the use of the ML method has three requirements: (a) a large sample size, (b) continuous levels of measurement, and (c) multivariate normality of the observed variables (Brown, 2006; Harrington, 2009). Fortunately, sample size was not an issue in the current research according to multiple “rules of thumb” discussed later to determine whether sample size is sufficient (Brown, 2006; Harrington, 2009). The second requirement states that observed variables should be measured on an interval or ratio scale of measurement. Unfortunately, many

social science researchers need to use surveys that are based on some Likert-type scale (e.g. very satisfied, somewhat satisfied). As discussed further in a subsequent section, these scales introduce possible biases due to violations of the normality assumption. As for the third requirement, the ML method is robust to small or moderate deviations from normality, especially with larger sample sizes (Brown, 2006). A benefit of using the ML method instead of other available methods (e.g., weighted least squares) is that it provides standard errors that allow for significance testing of parameter estimates and the construction of confidence intervals (Brown, 2006).

Evaluation of model fit. Once a model is specified based on previous research or theory, the statistical requirements above have been met, and the method for model estimation has been chosen, the model is evaluated for plausibility based on three sources of information: (a) model fit indices, (b) areas of poor fit (strain), and (c) parameter estimates.

Overall goodness of fit. There are multiple indices of model fit available when performing a confirmatory factor analysis; a review of all of them is beyond the scope of this study. What is important for the purposes of this study is the recommendation to examine multiple fit indices when determining whether a specified model has acceptable fit (Brown, 2006). Fit indices may be divided into three general (sometimes overlapping) categories: (a) indices of absolute fit, (b) indices that correct for model parsimony, and (c) indices of comparative fit. Brown (2006) suggested reporting at least one index of fit from each of these categories.

Following the guidelines of Brown (2006), multiple fit indices were reported in this research: χ^2 , RMSEA (Steiger & Lind, 1980), CFI (Bentler, 1990) and TLI (Tucker & Lewis,

1973). The classic fit index is χ^2 , an index of absolute fit. Statistically significant values for this index indicate that the hypothesized model does *not* fit the data well (i.e., that $S \neq \Sigma$). There are significant limitations to this fit index, however, that make researchers use additional fit indices to establish whether their models are plausible. First, χ^2 values are inflated for large sample sizes (thus, making it more difficult to deem a model as plausible even when the differences between S and Σ are minimal). Second, because it is an index of *absolute* fit, the χ^2 index is based on the conservative hypothesis that S is *exactly* equal to Σ .

A second category of fit indices includes those that adjust for model complexity. In other words, models that have fewer freely estimated parameters are considered better-fitting, compared to those with more degrees of freedom (df). The RMSEA, which indicates the extent to which a model fits *reasonably* well (as opposed to perfectly well), falls in this category of fit indices. This index is sensitive to the number of parameters in a model, but not sample size (as is χ^2).

The third category of fit indices, those that provide information of comparative fit, include CFI and TLI. Comparative fit indices are more liberal indices than absolute fit indices because they indicate the extent to which a model is better-fitting than a model with no relationships among its variables (instead of determining whether the model fits *exactly*). Finally, the TLI index is a “hybrid” index of comparative fit because it penalizes for model complexity.

For the current research, models with RMSEA values of .06 or below, and TLI and CFI indices greater than .95, were considered plausible models (Hu & Bentler, 1999). Models that do not meet these criteria might have one or more of the following problems: (a) too few factors, (b) too many factors, (c) inappropriate indicators, (d) an incorrect mapping of indicator variables

onto factors, or (e) inappropriate specifications of error correlations¹² (Brown, 2006; Harrington, 2009). For this reason, in addition to evaluating models by examining these fit indices, it was important to more closely evaluate the models for areas of poor fit (i.e., localized strain), and to examine parameter estimates (e.g., factor loadings) more closely.

Localized areas of strain. Overall goodness of fit indices provide only a part of the picture when evaluating the plausibility of a model. Sometimes, such indices suggest good fit but upon inspection, there are instances where the model does not fit the data well. Conversely, poor fit (as indicated by goodness of fit measures) does not provide the researcher with information regarding what, exactly, caused the poor fit. Two general methods are available for evaluating areas of poor fit: (a) analysis of residuals, and (b) modification indices. Standardized residuals indicate the difference between the sample variance-covariance matrix (S) and the hypothesized variance-covariance matrix (Σ). Because residuals are standardized, it is suggested that researchers scan the residuals to find values that are equal to or greater than the absolute value of 1.96 (a value corresponding to a significant z-score at the .05 alpha level) (Brown, 2006). A second way to evaluate areas of strain in a model is to examine the modification indices provided in the statistical software package. Modification indices are provided for fixed and constrained parameters, and provide an estimate of the decrease in χ^2 that would result from freely estimating the parameters.

Examination of parameters. Whereas an analysis of residuals and modification indices provide information about whether freely estimating a parameter might improve model fit, examining values for the parameters that are already freely-estimated provides additional

¹² Numerous or extremely large misspecifications should not occur, however, when the CFA model is based on prior empirical and conceptual evidence (Brown, 2006).

information about the fit of a model. Specifically, parameter values (e.g., factor loadings, correlations among factors) with z-scores above 1.96 are considered statistically significant. Those parameters that are not significant might be eliminated to create a better-fitting model.

Statistical significance does not, alone, indicate that a parameter is useful; the magnitude of parameter estimates should be evaluated to see if they are substantively meaningful (Brown, 2006). For example, factor loadings describe the extent to which indicator variables may be predicted from the latent variable; in this way, they are similar to regression coefficients in a regression analysis (Brown, 2006). The standardized factor loading for an indicator variable, when squared, represents its communality. *Communality* refers to the percent of variance an indicator shares with the latent variable with which it is associated. There is no standard rule to determine whether an indicator shares “enough” variance with the latent variable, but generally standardized factor loadings below .3 are considered poor and are not interpreted (Brown, 2006; Harrington, 2009). Another thing that should be inspected is the correlations among latent variables; factor correlations exceeding .80 might be indicative of poor discriminant validity (Brown, 2006).

Data Considerations

When conducting a confirmatory factor analysis, one must be mindful of the statistical assumptions and possible limitations underlying its use. Four issues are now considered in turn: sample size, missing data, normality, and outliers.

Sample Size and Missing Data

The sample size for this study ($n = 1,925$) exceeds the minimum guidelines provided throughout the literature (Brown, 2006; Kline, 2005; Tabachnick & Fidell, 2007). Among the

possible ways of dealing with missing data (e.g., listwise, pairwise), methodologists generally accept the Direct ML method to be the most appropriate within CFA (Allison, 2003; Brown, 2006). The use of Direct ML usually leads to the most consistent parameter estimates, standard errors, and test statistics (Brown, 2006). Other commonly used methods of handling missing data, such as listwise or pairwise deletion, may result in a loss of power and biased estimates, standard errors, and test statistics (Allison, 2003; Brown, 2006). Thus, for the current research, Direct ML within the AMOS program was used to estimate missing data.

Normality and Outliers

Multivariate normality¹³ was assessed using both SPSS/PASW and AMOS. SPSS was necessary because unlike AMOS, the program allows for the construction of histograms. Although univariate skewness and kurtosis estimates are provided by AMOS, this method is not as useful with large sample sizes because even very small deviations from normality may result in the conclusion that the distribution of a variable is not normal (Tabachnick & Fidell, 1997). In addition, with large sample sizes, univariate skewness and kurtosis do not affect findings as much as they would with smaller samples (Tabachnick & Fidell, 1997). Extreme non-normality (in particular, extreme kurtosis), however, has consequences that would warrant a change in the statistical analyses to be performed. Specifically, non-normality could result in an underestimation of the standard error, resulting in an inflated Type I error, incorrect parameter estimates, inflated tests of absolute fit, and underestimated tests of comparative fit (Brown, 2006). In this case, the unweighted least squares method, or UWLS, might be used instead of the

¹³ Although the assumption holds that multivariate normality is necessary, this is especially difficult to examine; thus, most methodologists find it acceptable to evaluate univariate normality to infer multivariate normality (Floyd & Widaman, 1995; Kline, 2005). One measure of multivariate normality, Mardia's test of normality (Mardia, 1974), is provided in AMOS output, but even high values may be associated with minimal bias (Muthén and Kaplan, 1985).

Maximum Likelihood method. Finally, outliers were detected using Mahalanobis d^2 ; using an iterative process, items with Mahalanobis d^2 values with $p < .001$ were deleted from the sample (Brown, 2006; Tabachnick & Fidell, 1997).

NSSE Benchmarks: Is the 5-Factor Model of Student Engagement Valid?

Internal Consistency

The internal consistency of the five NSSE benchmarks was evaluated, in part, by examining coefficient alphas. Because coefficient alphas are affected by the number of observed variables, and because these values can be high even when items are not highly correlated or when there are multiple dimensions, interitem correlations and item-total correlations were also computed (Cortina, 1993; Smith & McCarthy, 1995). Such information can provide more valid insights regarding the desired unidimensionality of each of the five benchmarks, especially for benchmarks having numerous observable variables (Briggs & Cheek, 1986; Clark & Watson, 1995). Items were expected to have high interitem and item-total correlations if they are representative of one (and only one) dimension (Bollen & Lennox, 1991). Items with low corrected item-total correlations ($< .40$) (Lounsbury, Gibson, & Saudargas, 2006) or low interitem correlations ($< .40$) with other items associated with the same benchmark were deemed statistically poor items. Similarly, benchmarks with low coefficient alphas ($< .75$) (Lounsbury et al., 2006) were considered to have poor internal consistency.

Confirmatory Factor Analysis

Model specification. The model evaluated using confirmatory factor analysis is shown in Figure 3.2.¹⁴ The five latent variables in this model represent the five benchmarks that are

¹⁴ This model varies from the original five-factor benchmark proposed by NSSE because some items were excluded from analyses (e.g., due to non-normality). Refer to the results section for a discussion of item deletion.

computed and reported by NSSE, whereas the observed variables are those items that comprise each benchmark. Because there is no theoretical reason to model correlated error variances, all error variances were fixed (i.e., correlations were not freely estimated and were set to equal 0). On the other hand, based on previous evidence that NSSE benchmarks are correlated, factor covariances were freely estimated.

To ensure that the model was identified, latent variables were scaled. To scale latent variables, the variances of the latent variables were all set to equal one.

Model estimation. The study's data met the requirements for the maximum likelihood (ML) method (e.g., sample size, normality). Consequently, this estimation procedure was used because it allows for the computation of standard errors and confidence intervals of parameter estimates (Brown, 2006).

Evaluation of model fit. The three general methods of evaluating a model's fit, as described previously, were used to evaluate the five-factor benchmark model of the NSSE instrument. Four measures of fit were used to evaluate the plausibility of the five-factor model: χ^2 , RMSEA, CFI, and TLI. As described earlier, these indices provide complementary types of information about the extent to which a model is plausible. Because overall goodness of fit indices do not provide specific information about areas of possible misspecification, modification indices were used to evaluate localized areas of strain. As a last step in evaluating

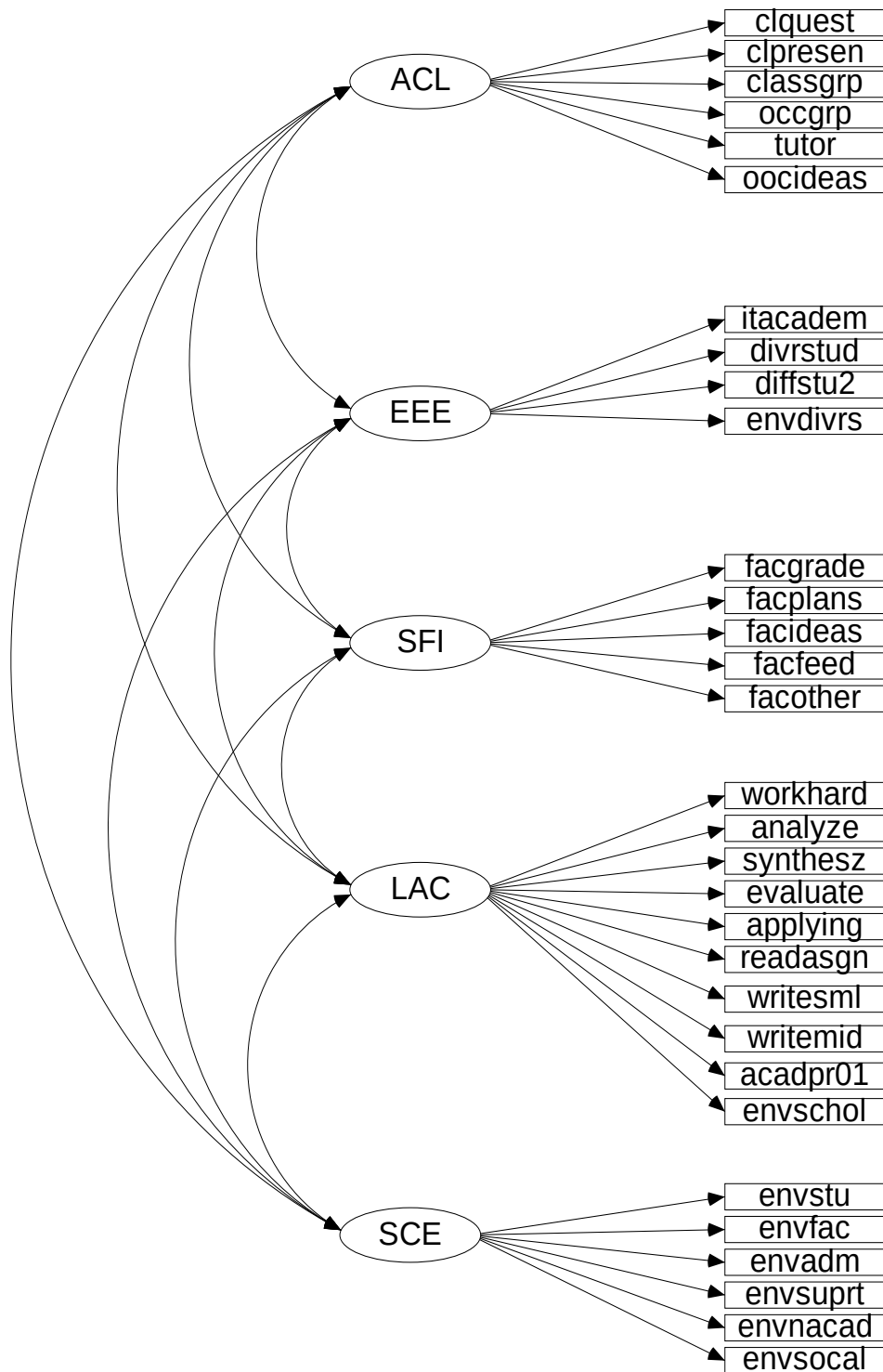


Figure 3.2. The five-factor NSSE benchmark model. Error associated with observed variables are not specified to correlate.

the benchmark model, freely estimated parameter estimates were examined. Observed variables with factor loadings of .3 were deemed “acceptable” (Brown, 2006). In addition, the correlations among the five latent benchmark factors were evaluated; those that were above .80 provided evidence against the discriminant validity of the individual benchmarks (Brown, 2006).

Pike’s Scalelets: Is the 12-Factor Model of Student Engagement Valid?

Internal Consistency

To evaluate the internal consistency of Pike’s (2006b) scalelet model, coefficient alphas, interitem correlations, and item-total correlations were computed. As with the benchmarks, statistically poor items were those with low interitem correlations ($< .40$) or low item-total correlations ($< .40$) (Lounsbury et al., 2006). In addition, benchmarks were considered to have low internal consistency if coefficient alpha was low ($< .75$) (Lounsbury et al., 2006).

Confirmatory Factor Analysis

Pike’s (2006b) scalelet model was evaluated using confirmatory factor analysis. The procedure used was exactly the same as that used when evaluating the NSSE benchmark model, except of course the specified model was different. The scalelet model (Pike, 2006b) is depicted in Figure 3.3.¹⁵ As in the benchmark model, latent variable variances were set to equal one to identify the model. Correlations among latent variables were freely estimated, and the correlations among the observed variables’ error variances were fixed to equal zero.

¹⁵ This model varies from Pike’s (2006b) proposed model because some items were excluded from the analyses. Refer to the results section for a discussion of the specific items (and scalelets) that were excluded.

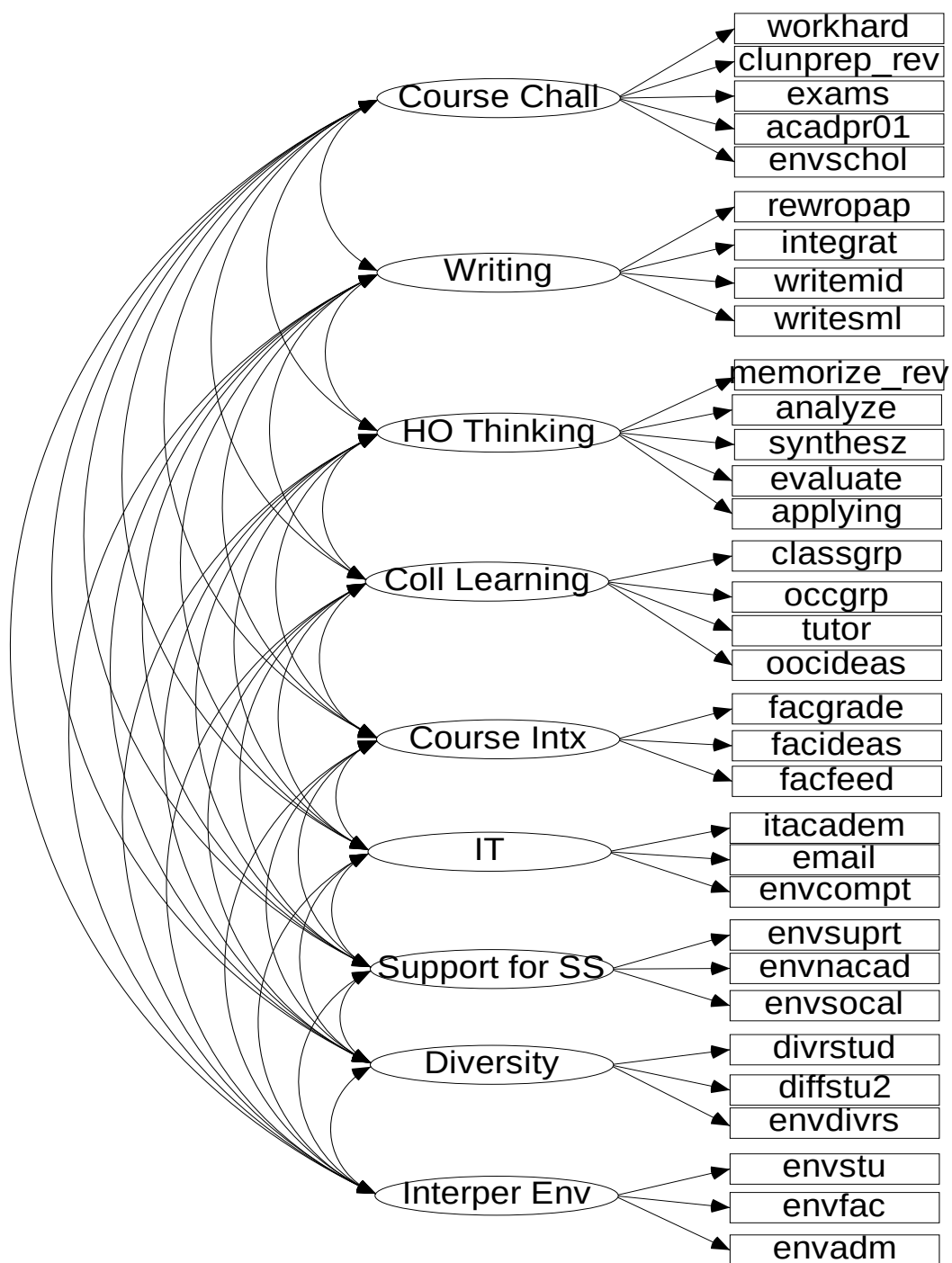


Figure 3.3. Pike's (2006b) scalelet model. Errors associated with observed variables are not specified to correlate.

Is There a Model That Better Fits the Data for The University of Tennessee, Knoxville?

Factor Analysis

The goal of exploratory factor analysis is to establish the underlying constructs, or factors, that explain the variation and covariation among multiple observed variables (Briggs & Cheek, 1986; Brown, 2006). Like confirmatory factor analysis, exploratory factor analysis is based on the common factor model which describes each measured variable as a linear function of two things: (a) common variance, and (b) unique variance (Brown, 2006; Thurstone, 1947). *Common variance* refers to the variance in a measured variable that is explained by a factor, whereas *unique variance* refers to systematic and random error variance in the measured variable (Floyd & Widaman, 1995). The primary difference between EFA and CFA is the significant a priori decisions (e.g., the number of factors, covariation among latent variables) that need to be made when using CFA (Brown, 2006). Exploratory factor analysis is, as its name implies, exploratory rather than confirmatory in nature. As such, it is a technique that is generally used in the beginning stages of construct validation (Brown, 2006). In fact, when EFA is used in the development of an instrument, one is more likely to find that the hypothesized model is a good fit when confirmatory factor analysis is applied to the data (Floyd & Widaman, 1995). The results from EFA and CFA may differ, however, because the two statistical techniques serve different purposes. The purpose of EFA is to create factors that account for a large proportion of the variance in a set of data, whereas the purpose of CFA is to assess the fit of a model based on the variance not accounted for by the factors (Floyd & Widaman, 1995). In other words, even after using EFA to create factors, confirmatory factor analysis may uncover some variance that is left unaccounted for by the factors (Floyd & Widaman, 1995).

Because neither the NSSE benchmark model nor Pike's (2006b) scalelet model appeared to have good model fit using CFA, factor analysis was used to create a model that best represented the data from the 2009 administration of the NSSE at The University of Tennessee, Knoxville. Principal axis factoring (PAF) was used to extract factors from the items included in the benchmarks or scalelets. Following factor extraction, a scree test (Cattell, 1966) in conjunction with an analysis of eigenvalues greater than 1, was used to select factors. Then, oblique rotation (i.e., Promax) was used to make the factors more interpretable. Items that were not correlated with any of the factors (i.e., factor loadings less than .3), as well as items that were crossloaded with two or more factors, were deleted (Brown, 2006; Schonrock-Adema et al., 2009). In addition, factors that were associated with variables all having low communalities, as well as those with only two variables associated with them, were deleted (Brown, 2006). Upon deleting individual items or factors as necessary, the steps above were repeated (e.g., extraction of factors, rotation) until an acceptable solution was found. Resulting factors were then evaluated for meaningfulness and interpretability.

Confirmatory Factor Analysis Using 2010 Data

The factor structure of a model established using exploratory factor analysis should be validated with CFA on a different sample (Brown, 2006; Haig, 2005). Thus, the factor solution obtained in the previous analysis was evaluated using a sample of data from the 2010 administration of the NSSE at The University of Tennessee, Knoxville. Consistency in results largely depended on the ability of the EFA solution to account for most of the variance in the data (Floyd & Widaman, 1995).

The procedure used to evaluate the resulting model was exactly the same as that used in previous confirmatory factor analyses in this study. To identify the model, latent variable variances were set to equal one. Correlations among latent variables were freely estimated, and the correlations among the observed variables' error variances were fixed to equal 0.

CHAPTER 4: RESULTS

The purpose of this chapter is to address, statistically, the three primary research questions of interest in this paper. First, the results from preliminary analyses of the data (i.e., checks for normality and outliers) are presented. Then, the statistical analyses used to address the three main research questions (i.e., those regarding the factorial validity of three different models of student engagement) are presented.

Normality and Outliers

Three NSSE items were particularly problematic because they exhibited both skewness and kurtosis based upon inspection of histograms: *writemor* (skew = 4.02, c.r. = 44.04; kurtosis = 18.48, c.r. = 101.28), *cocurr01* (skew = 1.57, c.r. = 24.38; kurtosis = 2.90, c.r. = 22.55), and *commproj* (skew = 1.52, c.r. = 23.65; kurtosis = 1.74, c.r. = 13.57) (see Figure 4.1 for histograms of these variables). These items were not included in any subsequent analyses. Although univariate skewness and kurtosis estimates should not be relied upon without also viewing histograms, especially with large sample sizes (Tabachnick & Fidell, 1997), these values are reported in Table A.3 for completeness.¹⁶

Outliers were dealt with by deleting those observations with Mahalanobis d^2 values of $p < .001$ (Brown, 2006; Tabachnick & Fidell, 1997). To accomplish this, it was necessary to request the “tests for normality and outliers” option in AMOS for each of the models depicted in Figures 3.2 and 3.3. For the NSSE benchmark model, the first iteration resulted in the deletion of 25 (less than one percent) students from the dataset. Because outliers may sometimes “hide behind” other outliers, a second iteration was performed and resulted in the deletion of eight

¹⁶ Critical ratio (c.r.) values should be interpreted as standardized deviation scores (z-scores). Using a two-tailed level of significance equal to .001, variables with skewness or kurtosis critical ratios equal to or greater than 3.30 would be considered non-normal.

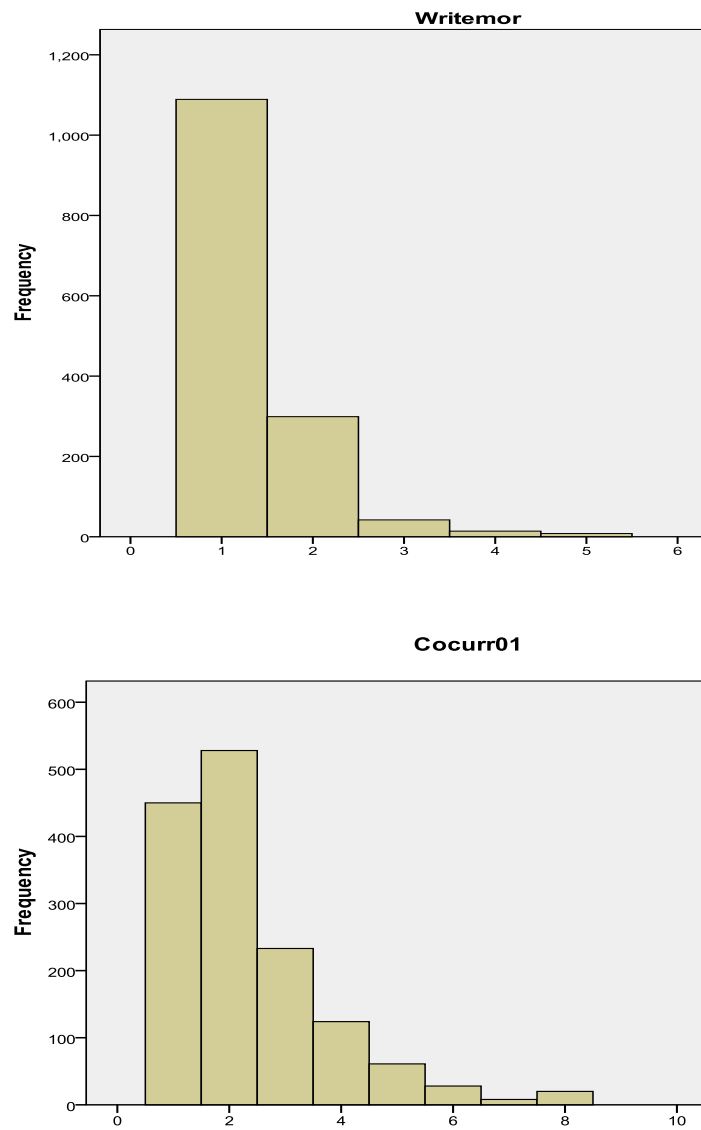
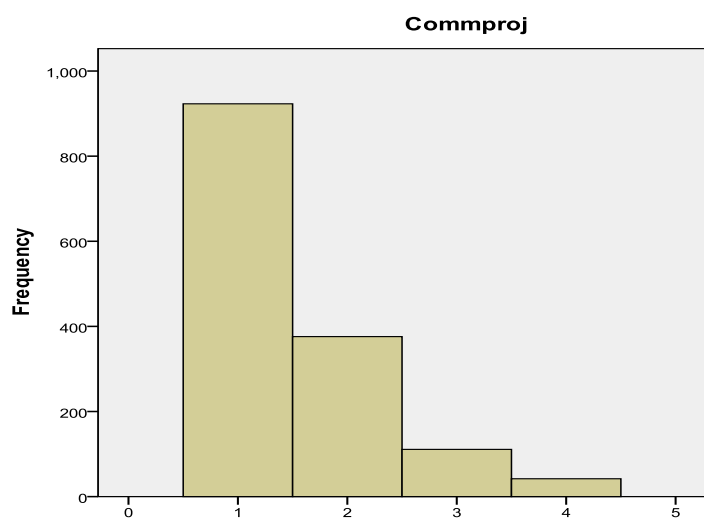


Figure 4.1. Normality: Histograms. The histograms of three variables, *writemor*, *cocurr01*, and *commproj*, were non-normal and were omitted from all analyses.

Figure 4.1 (continued)



more students. The final dataset used to evaluate the benchmark model consisted of 1892 students (958 first-year students, and 934 seniors). For Pike's (2006b) scalelet model, the first iteration resulted in the deletion of 32 (1.66%) students, and the second iteration resulted in the deletion of nine more cases. The final dataset used to evaluate Pike's (2006b) scalelet model consisted of NSSE data collected from 1884 students (963 first-year students, and 921 seniors).

NSSE Benchmarks: Is the Five-Factor Model of Student Engagement Valid?

Internal Consistency

Multiple measures of the internal consistency of the five NSSE benchmarks are reported in Table A.4. Specifically, coefficient alphas for each of the five benchmarks are provided, as well as interitem correlations and item-total correlations for each benchmark item. Using $\alpha = .75$ as a rule of thumb for acceptable coefficient alpha (e.g., Lounsbury et al., 2006), only one benchmark (*SFI*, $\alpha = .75$) exhibited sufficient internal consistency. Two other benchmarks nearly met this criterion (*LAC*, $\alpha = .71$, and *SCE*, $\alpha = .73$). The other benchmarks exhibited poor internal consistency based on coefficient alpha ($\alpha = .62$ and $.57$ for *ACL*, and *EEE*¹⁷, respectively).

Because coefficient alpha values provide only a piece of the picture regarding a scale's internal consistency (Briggs & Cheek, 1986; Clark & Watson, 1995; Cortina, 1993; Smith & McCarthy, 1995), interitem correlations and item-total correlations were also computed and are reported in Table A.4. Of 91 interitem correlations, only 15 (16.48%) met or exceeded the criterion of $r = .40$ (e.g., Lounsbury et al., 2006). Using the same criterion of $r = .40$, corrected

¹⁷ The Enriching Educational Experiences (EEE) benchmark was reduced from 12 items to 4 items because of the nominal scale of measurement of several items, and because of the removal of cocurr01 due to non-normality. Because α increases as the number of scale items increases (Cortina, 1993; Smith & McCarthy, 1995), an analysis including all 12 original benchmark items would likely be higher.

item-total correlations were evaluated for each benchmark item. These values provide information about how well an item is correlated with its benchmark (with that item removed). All items ($n = 11$) in the *Student-Faculty Interaction (SFI)* and *Supportive Campus Environments (SCE)* exhibited sufficient correlation with their corresponding benchmarks. Nearly half of the items (8 out of 20, or 40%) comprising the remaining benchmarks exhibited sufficient correlation with their corresponding benchmarks. On the other hand, some items had very low correlations with their corresponding benchmarks (e.g., the item-total correlation for *envdivrs* in the *Enriching Education Experiences* benchmark was .17).

To summarize, some NSSE benchmarks were more internally consistent than others. The *Student-Faculty Interaction (SFI)* and *Supportive Campus Environments (SCE)* benchmarks were adequate in terms of internal consistency, whereas the other three benchmarks did not appear to be comprised of items that were entirely unidimensional in content.

Confirmatory Factor Analysis

A three-step process was used to evaluate the measurement model of the five-factor NSSE benchmark model. First, four complementary measures of fit were used to evaluate the plausibility of the five-factor model: χ^2 , RMSEA, CFI, and TLI. Second, modification indices were used to evaluate localized areas of strain. Third and lastly, freely estimated parameter estimates were examined.

Overall goodness of fit. Following the guidelines of Brown (2006), multiple fit indices were computed. Using the criteria suggested by Hu and Bentler (1999), a model is considered plausible if RMSEA is .06 or below, and TLI and CFI indices are both greater than .95. Based on these criteria, the five-factor measurement model of the NSSE did not appear to be a good-fitting

model ($\chi^2 = 4066.80$, $p < .001$; RMSEA = .07; CFI = .77; and TLI = .74). The next step, then, was to evaluate the model for areas of poor fit (i.e., localized strain).

Localized areas of strain. An analysis of residuals and an analysis of modification indices was used to evaluate the areas of poor fit in the benchmark model. Standardized residuals indicated the difference between the sample variance-covariance matrix (S) and the hypothesized variance-covariance matrix (Σ). Items with values exceeding $|1.96|$ ($p > .05$) were deemed to have an excessive amount of residual. Table A.5 displays the number of residuals exceeding the criterion value of $|2|$ (1.96 being rounded slightly for simplicity) for each NSSE benchmark item. Those NSSE items with the fewest standardized residuals (≤ 5) exceeding the criterion value of $|2|$ also tended to have high item-score correlations. For example, the item *synthesz*, which had only five residuals above $|2|$, had an item-total correlation equal to .56. Unexpectedly, items with the most (≥ 15) standardized residuals in excess of $|2|$ also tended to have acceptable (or nearly acceptable) item-total correlations, with the exception of *envdivrs*. *Envdivrs*, which had the lowest item-total correlation of .17, had both the *most* and the *largest* standardized residuals. For instance, the standardized difference between the sample and hypothesized matrices for the correlation between *envdivrs* and *envsuprt* was 18.0. That there was so much residual, and that *envdivrs* had such a low item-total correlation, suggests that the item is problematic.

Another way of evaluating areas of strain in the five-factor NSSE benchmark model was to examine the modification indices provided in AMOS. Modification indices provide an estimate of the decrease in χ^2 that would result from freely estimating parameters. In Table A.6,

there are two columns. The column on the left displays the parameters to be freely estimated, and the column on the right shows the corresponding modification indices.

Upon inspection, several modification indices in Table A.6 are worth discussing. First, freely estimating (i.e., adding a causal path between) the residuals associated with (a) *envdivrs* and *envsocal*, (b) *envdivrs* and *envsuprt*, and (c) *envdivrs* and *envnacad* would lead to model improvement. This finding was likely due to the fact that the content of *envdivrs* is so similar to that of the *SCE* items. Specifically, *envdivrs* and the items comprising the *SCE* benchmark all ask students about their campus environment and relationships with others (refer back to Table A.1 for a description of NSSE benchmark items). In fact, results also suggest that the measurement model would be improved upon by associating *envdivrs* with the *SCE* factor (MI = 387.10).

Additionally, modification indices suggested freely estimating the relationships between several variables that were included within the same benchmark. For instance, the modification index of 218.6 associated with freely estimating the residuals between *envfac* and *envadm* suggested that these two variables had something in common (other than their shared underlying construct, *SCE*). The wording of these items is almost identical. The only difference between these items is that one asks students about their relationships with faculty, whereas the other asks them about their relationships with administrative personnel. As another example, the modification index of 172.4 associated with freely estimating the relationship between the residuals of *writemid* and *writesml* indicated that students' responses to these items were associated in ways that were not due to them both being associated with LAC. The wording of

these two items is highly similar, and both items deal with the extent to which students engage in writing.

Examination of parameters. Standardized factor loadings and communalities for items in the five-factor benchmark model are displayed in Table A.7. Four items had standardized loadings below .3 and were thus considered poor (Brown, 2006; Harrington, 2009): *envdivrs*, *readasgn*, *itacedem*, and *writesml*. Removal of these items from the model would result in a better-fitting model, except for *envdivrs*. Based on its modification index, retaining it in the model, but including it as an item in the *SCE* benchmark, would probably be a more suitable solution. Finally, the correlations among latent variables were inspected because correlations exceeding .80 are indicative of poor discriminant validity (Brown, 2006). The correlation between the *ACL* and *SFI* benchmarks was equal to .90, providing evidence against the discriminant validity of these two benchmarks.

Summary of Benchmark Results

Altogether, results suggest that the five-factor benchmark model lacks validity. At an item level of investigation, only 19 of 35 items had sufficient item-total correlations. All the items in the *SFI* and *SCE* benchmarks were correlated to a sufficient degree to their benchmarks. On the other hand, most of the items in the *ACL* benchmark were not correlated substantially with the overall benchmark. In the residual analysis, four items appeared especially problematic: *facother* (*SFI*), *workhard* (*LAC*), *envstu* (*SCE*), and *envdivrs* (*EEE*). Results from the confirmatory factor analysis also pointed to *envdivrs* as a problematic item; results indicated that the item might belong in the *SCE* benchmark. In addition, *envdivrs*, *readasgn*, *itacadem*, and *writesml* were not sufficiently related to their benchmark according to the parameter estimates.

Moving up to the benchmark level of analysis, two of the five benchmarks (i.e., *ACL* and *EEE*) exhibited poor internal consistency as measured by coefficient alpha. Also problematic was the finding in the confirmatory factor analysis that the *SFI* benchmark was highly correlated with the *ACL* benchmark. This suggested a lack of discriminant validity. At the largest level of examination, the model level, results suggested a discrepancy in the model and the data. In other words, confirmatory factor analysis results indicated that the five-factor benchmark model is invalid as a model of student engagement for the UT population of students.

Pike's Scalelets: Is the 12-Factor Model of Student Engagement Valid?

Internal Consistency

Coefficient alphas for the nine scalelets (Pike, 2006b) evaluated in the current study, as well as interitem correlations and item-total correlations, are shown in Table A.8. None of the scalelets reached the criterion coefficient alpha level of .75 suggested in the literature (e.g., Lounsbury et al., 2006), although the *Support for Student Services* scalelet was close ($\alpha = .73$). Coefficient alphas for the remaining eight scales ranged from .49 (*Instructional Technology*) to .68 (*Higher Order Thinking* and *Collaborative Learning*). Of 62 interitem correlations, 13 (21%) met or exceeded the criterion of $r = .40$ (e.g., Lounsbury et al., 2006). Using the same criterion of $r = .40$, corrected item-total correlations were evaluated for each of the 36 scalelet items. Of the 36 items, 19 (53%) had item-total correlations reaching or exceeding $r = .40$. None of the items comprising the *Instructional Technology* scalelet or the *Course Challenge* scalelet had item-total correlations meeting this criterion level. On the other hand, the items comprising the *Course Interaction* scalelet, as well as the *Support for Student Services* scalelet, all reached the criterion item-total correlation of .40. For the *Higher Order Thinking* scalelet, only one of five items

failed to reach the criterion (*memorize_rev*). Similar to the conclusion reached when evaluating the internal consistency measures of the benchmarks, it appears that some scalelets (e.g., *Support for Student Services*) are more unidimensional than others (e.g., *Course Challenge*).

Confirmatory Factor Analysis

A three-step process was used to evaluate Pike's (2006b) model of student engagement. First, four measures of fit were used to evaluate the measurement model: χ^2 , RMSEA, CFI, and TLI. Second, residuals and modification indices were used to evaluate localized areas of strain. Finally, freely estimated parameter estimates were examined.

Overall goodness of fit. Following the guidelines of Brown (2006), multiple fit indices were computed. As previously discussed, a model is considered plausible if RMSEA is .06 or below, and TLI and CFI indices are both greater than .95 (Hu & Bentler, 1999). Based on these criteria, the scalelet model of the NSSE (Pike, 2006b) did not appear to be a good-fitting model ($\chi^2 = 3176.81$, $p < .001$; RMSEA = .06; CFI = .82; and TLI = .80). The next step was to evaluate the model for areas of poor fit (i.e., localized strain).

Localized areas of strain. Two methods were used to evaluate poorly-fitting areas of the measurement model: (a) analysis of residuals, and (b) modification indices. Standardized residuals indicated the difference (i.e., residual) between the sample variance-covariance matrix (S) and the hypothesized variance-covariance matrix (Σ). Items with values exceeding $|1.96|$ ($p > .05$) were deemed to have an excessive amount of residual. Table A.9 displays the number of residuals exceeding the criterion value of $|2|$ ($|1.96|$ was rounded for simplicity) for each of the scalelet items. Five scalelet items had relatively few (≤ 5) standardized residuals that exceeded the 2.0 ($p > .05$) cutoff: *divrstud*, *itacadem*, *oocideas*, *synthesz*, and *integrat*. As in the analysis of

residuals for the benchmark model, the item *envdivrs* had the most ($n = 23$) standardized residuals. That there was so much residual, and that *envdivrs* had such a low item-total correlation ($r = .17$), suggests that the item is problematic in Pike's (2006b) scalelet model.

Another way of evaluating poorly-fitting areas in the measurement model was to examine the modification indices provided in AMOS. Table A.10 displays the modification indices that would result by freely estimating (i.e., adding a line between) pairs of variables. Results suggested that freely estimating (i.e., adding a line between) the residuals of (a) *facgrade* and *email*, (b), *envschol* and *envsuprt*, and (c) *writemid* and *writesml* would result in significant decreases in χ^2 . The modification indices for pairs a and b are particularly difficult to explain because the variables in each pair belong to different scalelets. For instance *facgrade* is an item in the *Course Interaction* scalelet, whereas *email* belongs to the *Information Technology* scalelet. The residuals for these items might be correlated because students used email to discuss grades with faculty members. Results also suggest that *envdivrs*, as modeled, is problematic. It is supposed to be associated with the *Diversity* scalelet, and *only* the *Diversity* scalelet. However, modification indices indicated that associating the variable with three other scalelets (i.e., *Support for Student Success*, *Course Challenge*, and *Interpersonal Environment*) would lead to model improvement.

Examination of parameters. Standardized factor loadings and communalities for Pike's (2006b) scalelet model are displayed in Table A.11. Four items had loadings below .3 and were thus considered poor (Brown, 2006; Harrington, 2009): *envdivrs*, *writesml*, *clunprep_rev*, and *memorize_rev*. Removal of these items from the model results in a better-fitting model. Next, the correlations among latent variables were inspected for evidence of poor discriminant validity

(Brown, 2006). The correlations between the *Course Interaction* scalelet was highly correlated with two other scalelets: *Collaborative Learning* ($r = .84$) and *Information Technology* ($r = .83$).

Summary of Scalelet Results

Similar to the results of the benchmark model, the scalelet model of student engagement generally lacked validity. At an item level of investigation, only 19 of 36 items had sufficient item-total correlations. In the residual analysis, two items appeared especially problematic: *envdivrs* (*Diversity* scalelet) and *envstu* (*Interpersonal Environment* scalelet). Results from the confirmatory factor analysis also pointed to *envdivrs* as a problematic item; the item was erroneously associated with three scalelets to which it was hypothesized to be uncorrelated. In addition, *envdivrs*, along with three other variables (i.e., *clunprep_rev*, *writesml*, and *memorize_rev*), were not sufficiently related to their scalelets based on their small parameter estimates.

At the scalelet level of analysis, only one of the scalelets, *Support for Student Services*, exhibited (nearly) sufficient internal validity as measured by coefficient alpha. Also problematic was the finding in the confirmatory factor analysis that the *Course Interaction* scalelet was highly correlated with the *Collaborative Learning* and *Information Technology* scalelets. This suggested a lack of discriminant validity. At the largest level of examination, the model level, confirmatory factor analysis results indicated that the scalelet model, like the benchmark model, is invalid as a model of student engagement for the population of UT students.

Is There a Model That Better Fits the Data for The University of Tennessee, Knoxville?

Factor Analysis

First iteration. Because confirmatory factor analysis results suggested that neither the NSSE benchmark model nor the modified scalelet model (Pike, 2006b) was a valid model of student engagement, factor analysis was used to create a better model that represented the 2009 NSSE data collected at The University of Tennessee, Knoxville. Principal axis factoring (PAF) with oblique rotation (Promax) was used to extract factors from the items included in the benchmarks or scalelets. Following factor extraction, a scree test (Cattell, 1966; see Figure 4.2) in conjunction with an analysis of eigenvalues greater than one, resulted in 10 factors that cumulatively explained approximately 41% of the variance (see Table A.12 for the resulting pattern matrix).¹⁸ Six items failed to load on any of the 10 factors: *facfeed*, *clquest*, *clunprep_rev*, *itacadem*, *envcompt*, and *memorize_rev*. Additionally, *envstu* cross-loaded on two factors. Importantly, two factors (factor 6 and factor 10) were each associated with only two items. This suggested a better solution might have fewer factors.

Subsequent iterations. Once weaker items were deleted (i.e., those with only low factor loadings, and the one that crossloaded on two factors), subsequent iterations were done in order to arrive at an acceptable factor solution (i.e., one in which all variables loaded onto one and only one factor, and one in which all factors had at least three items loaded onto it). As in the first iteration, Principal axis factoring (PAF) with oblique rotation (Promax) was used to extract factors from benchmark and scalelet items in each of the subsequent iterations. The second factor analysis yielded nine factors that explained approximately 43% of the variance; however, two

¹⁸ The scree plot method was less useful in this study because the point at which “the elbow bends” is not clear. Thus, factors with eigenvalues greater than one were retained.

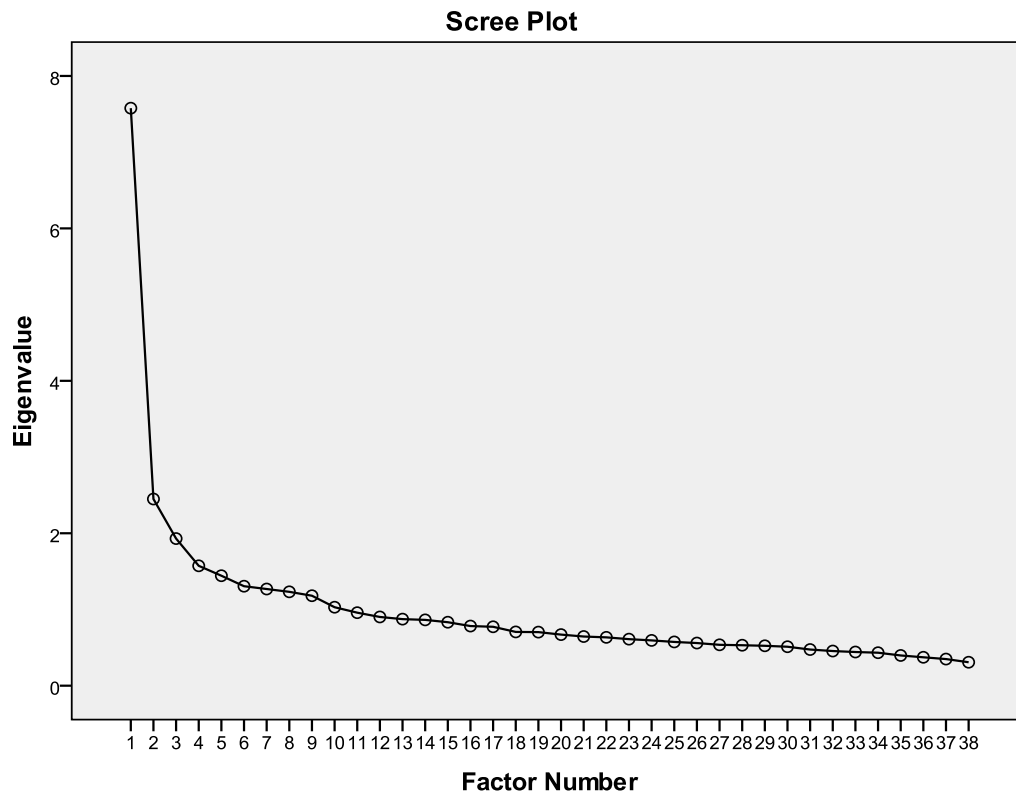


Figure 4.2. Factor analysis: Scree plot. This represents the scree plot of all NSSE items included in the benchmark or scalelet models. When the scree plot method is used to decide on the number of factors in a factor analysis, the bend (or the “elbow”) represents the suggested cut-off point.

items failed to load on any of the nine factors: *rewropap*, and *integrat*. Also, the item *facplans* cross-loaded on two factors. It was also discovered that the ninth factor had only two variables associated with it, *envfac* and *envadm*. For this reason, a third factor analysis was conducted. In this iteration, problematic items from the second iteration were removed and the variables were forced to converge into eight factors. These eight factors altogether explained about 43% of the variance, but one item, *classgrp*, failed to load on any factors. Also, two factors consisted of only two items each. Thus, a fourth iteration excluding *classgrp* was conducted in which the items were forced to converge into six factors that accounted for approximately 39 percent of the variance in the correlation matrix. This iteration resulted in an acceptable factor solution. Factors were then evaluated based on meaningfulness and interpretability, and factors were named. The final factor solution is displayed in Table A.13.

Confirmatory Factor Analysis

Confirmatory factor analysis was applied to a different NSSE dataset (i.e., from the 2010 administration at The University of Tennessee, Knoxville, $n = 2114$) to evaluate the six-factor solution. The same three-step process that was used in previous confirmatory factor analyses was used to evaluate the plausibility of the six-factor model.

Overall goodness of fit. Using Hu and Bentler's (1999) criteria, the six-factor measurement model of the NSSE did not appear to be a good-fitting model ($\chi^2 = 2913.94$, $p < .001$; RMSEA = .06; CFI = .83; and TLI = .80).

Localized areas of strain. Two methods were used to evaluate the areas of poor fit in the measurement model: (a) analysis of residuals, and (b) modification indices. Table A.14 displays the number of residuals exceeding the criterion value of $|2|$ for each of the items. Four items had

relatively few (≤ 5) standardized residuals that exceeded the $|2|$ ($p > .05$) cutoff: *diffstu2*, *analyze*, *envsuprt*, and *writemid*. The survey items *oocideas* and *envfac* had the most ($n = 21$ and 19 , respectively) standardized residuals.

Another way of evaluating areas of poor fit in the six-factor measurement model was to examine the modification indices provided in AMOS (see Table A.15). Results suggested that freely estimating (i.e., adding a line between) the residuals of (a) *envnacad* and *envsocal*, (b), *envfac* and *envadm*, (c) *email* and *facgrade*, (d) *clpresen* and *occgrp*, and (e) *envschol* and *envsuprt* would result in significant decreases in χ^2 . The modification index for the relationship between the errors associated with *envschol* and *envsuprt* is probably due to similar wording of the two items, as both items refer to academics. Additionally, results suggested that associating *workhard* with the *Academic Scholarship* factor would result in model improvement.

Examination of parameters. Standardized factor loadings and communalities for the six-factor model are displayed in Table A.16. None of the items had standardized loadings below .3. This was somewhat expected, because factor analysis was used to create the factors in the first place. The correlations among latent variables were inspected for evidence of poor discriminant validity (Brown, 2006). Correlations among factors ranged from low to moderate (i.e., $r = .04-.60$); none of the correlations were high enough to warrant concern about discriminant validity.

Summary of Factor Analysis and CFA of Six-Factor Model

The six-factor model of student engagement, although derived using a solid statistical technique (i.e., factor analysis), was found to be unacceptable when confirmatory factor analysis was applied on a different NSSE dataset. As previously discussed, confirmatory factor analysis is

more likely to indicate that a model is plausible when EFA is used to initially establish factors. This is not always the case, however, because of differences between the two statistical techniques. Exploratory factor analysis is used to create factors that account for a large proportion of variance in the data, whereas the purpose of *confirmatory* factor analysis is to assess the fit of a model based on the variance *not accounted for* by the factors (Floyd & Widaman, 1995). In the case of the six-factor model of student engagement, CFA indicated that a substantial amount of variance was left unaccounted for by the six factors derived from the exploratory factor analysis.

Overall Summary of Results

In short, none of the NSSE models (i.e., benchmark, scalelet, six-factor models) were found to have adequate factorial validity in the current study when applied to sample data from The University of Tennessee, Knoxville. This means that overall, NSSE survey items were not sufficiently associated with factors in any of the three models explored. According to Floyd and Widaman (1995), "...the failure to obtain acceptable confirmation using CFA indicates the scale[s] [require] modification to improve the measurement precision of [their] items" (p. 296).

CHAPTER 5: DISCUSSION

The purpose of this chapter is threefold: (a) to provide a summary of findings from the current study and relate these findings to those found in other research studies, (b) to discuss findings in greater detail, and (c) to discuss validity as an ongoing process by providing ideas for future research on the NSSE.

Summary of Findings

This study was designed to examine the factorial validity of three representations of the National Survey of Student Engagement: (a) the benchmark model, established and widely reported by the NSSE Institute; (b) the scalelet model, created by Pike (2006b); and (c) a six-factor model based on the results of a factor analysis. Results from this study's analyses generally suggested a lack of factorial validity for both the benchmark model and Pike's (2006b) scalelet model of the NSSE. When an exploratory factor analysis was used to produce another model of the NSSE that might be more appropriate for the data at The University of Tennessee, Knoxville, a confirmatory factor analysis failed to support the validity of the resulting six-factor model.

Similar to the findings of LaNasa et al. (2009) and Swerdzewski et al. (2007), the benchmark model used by the NSSE Institute to describe the NSSE failed to demonstrate sufficient factorial validity. As in the studies by LaNasa et al. (2009) and Swerdzewski et al. (2007), overall fit indices in this study suggested less than adequate model fit for the benchmark model of the NSSE. A comparison of parameter estimates yielded in this study and in LaNasa et al.'s (2009) study is informative.¹⁹ Five items had particularly high loadings ($\geq .70$) onto their

¹⁹ Parameter estimates were not reported in Swerdzewski et al.'s (2007) study.

benchmarks in both studies: *divrstud*, *diffstu2*, *analyze*, and *synthesize*. LaNasa et al. (2009) found six other items that loaded onto their benchmarks sufficiently: *envfac*, *envsuprt*, *envnacad*, *envsocal*, *applying*, and *evaluate*. Although these six items did not have standardized factor loadings greater than .70 in the current study, all six of them had factor loadings of at least .60. For instance, the factor loading for *evaluate* was .73 in LaNasa et al.'s (2009) study. In the current study, the standardized parameter estimate for *evaluate* was .69. On the other hand, two items (*readasgn* and *itacedem*) were particularly problematic (i.e., had factor loadings less than .30) in both studies.

LaNasa et al. (2009) did not report modification indices in their study of the NSSE benchmarks, but they did report correlations among NSSE benchmarks. A high correlation emerged between the *ACL* and *SFI* benchmarks in both this study and in LaNasa et al.'s (2009) research [$r = .89$ in LaNasa et al.'s (2009) study, $r = .90$ in this study]. To summarize, results from this study and LaNasa et al.'s (2009) research suggest that although some NSSE items are associated with NSSE benchmarks to a sufficient degree (e.g., *divrstud*), most items are not sufficiently representative of the overarching constructs they are supposed to measure. A high correlation between two benchmarks (*ACL* and *SFI*), along with poor overall fit indices, suggest that the NSSE items do not group together into five benchmarks as intended.

The factorial validity of another model of the NSSE, a modified version of Pike's (2006b) scalelet model, was also investigated in this study.²⁰ Simply put, the model exhibited poor overall fit. Four items had standardized loadings below .30: *envdivrs*, *writesml*, *clunprep_rev*, and *memorize_rev*. In addition, the *Course Interaction* scalelet was highly

²⁰ Recall that the scalelet model of the NSSE could not be investigated exactly as depicted by Pike (2006b) due to non-normality or nominal response options.

correlated with two scalelets: *Collaborative Learning* and *Information Technology*. As this was the first study to investigate the factorial validity of the Pike's (2006b) scalelets, further studies are needed to see if these results would replicate in other student populations.

Finally, a factor analysis was conducted to see if a better-fitting model could be produced from the 38 items comprising the NSSE benchmarks or scalelets. A six-factor model accounted for approximately 39 percent of the variance. This model, although it was created using exploratory factor analysis, was not shown to be factorially valid when it was applied to a different dataset using confirmatory factor analysis. In summary, none of the models explored (i.e., the benchmark model, the scalelet model, or the 6-factor model) was a valid representation of student engagement.

Discussion

Why did all three models tested in this study exhibit less than adequate fit? A related question might be, Does the NSSE instrument effectively capture the construct of student engagement for *any* students? To begin to answer these questions, one must consider the methods employed in the current study. The sample of students from The University of Tennessee, Knoxville, may not be representative of other college students. As such, results produced from other student samples might lend evidence in support of the factorial validity of the NSSE. For the benchmark model analyses, however, many of the results were similar to those of LaNasa et al. (2009) and Swerdzewski et al. (2007). It is important to replicate this study in various populations of students, as this was the first study to explore the factorial validity of Pike's (2006b) scalelet model of student engagement. Another issue in this study was the inability to include several NSSE items in analyses due to non-normality or nominal scales of

measurement. As a result, Pike's (2006b) model could not be evaluated exactly as created because the elimination of problematic items deemed it necessary to also reduce the number of scalelets. Had all items been appropriate for inclusion in the analysis, the model might have exhibited adequate fit. In the future, researchers asking whether the NSSE scalelet model is factorially valid might consider using statistical procedures that carry with them fewer statistical assumptions such as normality. Finally, in the current study, data from first-year students and seniors were combined. This could result in the confounding issue of student class rank. For example, it might be that the benchmark model is valid for first-year students, but not for seniors. As another example, the scalelet model might be an appropriate depiction of student engagement for seniors, but not for first-year students. Further studies that take student class rank into account are necessary.

If future studies—ones that can account for these kinds of issues (e.g., differences in student populations, the effect of class rank)—continue to find the benchmark model and Pike's (2006b) scalelet model lacking in terms of factorial validity, the items comprising the NSSE instrument might be to blame. As discussed at length earlier, the manner by which the NSSE was constructed in the first place (i.e., using a blend of theory and empirical analysis; see Kuh, 2001a and Kuh et al., 2001) brings about many questions regarding the instrument's validity. For instance, empirical procedures for creating NSSE benchmarks included both exploratory factor analysis *and* principal components analysis, a practice strongly suggesting a weak theoretical foundation for student engagement subscales and the items comprising them.

Future studies might also address validity evidence that is lacking in the NSSE literature. For example, a study investigating the discriminative validity of the NSSE benchmarks would

involve comparing NSSE results for groups known to differ, or presumed to differ, in student engagement (based on some other measure) (Benson & Clark, 1983; Messick, 1995). The predictive and incremental validity (see Koeske, 1994) of the NSSE have yet to be sufficiently explored, as well. It might be the case, for instance, that NSSE results are no more predictive of student outcomes than student demographic information such as socioeconomic status. Finally, the consequential validity of the NSSE (depicted by the benchmarks, scalelets, or any other means) needs to be examined. That is, what are the intended and unintended consequences of administering the NSSE and using its results to make decisions?

Should institutions such as The University of Tennessee, Knoxville continue to use the NSSE? Simply stated, more research is needed to answer this question. Results from this study (as well as those by LaNasa et al., 2009 and Swerdzewski et al., 2007) certainly suggest that institutional researchers and other stakeholders in higher education should avoid blindly accepting the NSSE instrument as a valid measure of student engagement. Validation is an ongoing process, as Messick (1993) described: "...[E]xisting validity evidence becomes enhanced (or contravened) by new findings...Inevitably, then, validity is an evolving property and validation is a continuing process" (p. 13). Unfortunately, however, researchers often believe that initial validity evidence presented on an instrument (for instance, in a technical manual) is sufficient to warrant the use of an instrument indefinitely. Smith and McCarthy (1995) wrote, "It is far too often the case that, once items are first written, their content validity appears taken as a given and is never scrutinized again" (p. 305). Similarly, Benson and Clark (1982) stated that "[t]he validation of a newly developed instrument is almost never accomplished through one study or by one researcher" (p. 798).

Despite the overwhelming popularity and reliance on the NSSE, its validity needs to be continually investigated. Without such studies, “inferences made from instruments that are even as well-known as NSSE must be made with extreme vigilance” (Swerdzewski et al., 2007, p. 18). The widespread use of the National Survey of Student Engagement might be useful for institutions wanting to compare their students to those at other institutions. Institutions might also be able to get a sense of intra-institutional changes with NSSE data. On the other hand, if various compositions of NSSE items into subscales (e.g., benchmarks or scalelets) are not valid in the first place, one is left with the question, What information is actually gained by making such comparisons? For example, what does it mean if a particular benchmark score at one institution is higher than that at another institution? The answers to these types of questions, without sufficient evidence for the NSSE’s factorial validity, remain unknown.

LIST OF REFERENCES

- Abdi, H. (2007). Bonferroni Test. In N. Salkind (Ed.), *Encyclopedia of Measurement and Statistics* (Vol. 1). Thousand Oaks: Sage.
- Ahren, C., Ryan, H. G., & Massa-McKinley, R. (2008). Assessment matters: The why and how of cracking open and using assessment results. *About Campus*, 13, 29-32.
- Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology*, 112, 545-557.
- American Educational Research Association/American Psychological Association/National Council on Measurement in Education (AERA/APA/NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- American Psychological Association/American Educational Research Association/National Council on Measurements Used in Education. (1955). Technical recommendations for psychological tests and diagnostic techniques. Washington, DC: American Psychological Association.
- Anastasi, A., & Urbina, S. (1997). *Psychological Testing* (Vol. 7). Upper Saddle River, NJ: Prentice Hall.
- Astin, A.W. (1993). *What matters in college? Four critical years revisited*. San Francisco: Jossey-Bass.
- Banta, T. W., Pike, G. R., & Hansen, M. J. (2009). The use of engagement data in accreditation, planning, and assessment. *New Directions for Institutional Research*, 2009(141), 21-34.

- Barnett, D.W., Lentz, F.E., & MacMann, G. (2000). In Behavioral assessment in schools: Theory, research, and clinical foundations. 2nd edition. In E. S. Shapiro & T. R. Kratochwill (Eds.)
- Benson, J., & Clark, F. (1982). A guide for instrument development and validation. *The American Journal of Occupational Therapy*, 36, 789-800.
- Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin*, 107, 238-246.
- Boehnke, M. (2010). UT students say they're less satisfied than peers. *Knoxville News Sentinel*.
- Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, 110, 305-314.
- Bowman, N. A. (2010). Can 1st-year college students accurately report their learning and development? *American Educational Research Journal*, 47, 466-496.
- Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality*, 54, 106-148.
- Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: The Guilford Press.
- Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (2nd ed.). New York: Routledge.
- 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.

- Carle, A. C., Jaffee, D., Vaughan, N. W., & Eder, D. (2009). Psychometric properties of three new National Survey of Student Engagement based engagement scales: An item response theory analysis. *Research in Higher Education*, 50, 775-794.
- Cattell, R. B. (1965). *The scientific analysis of personality*. Baltimore: Penguin Books, Inc.
- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245-276.
- Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 17, 309-319.
- Comrey, A. L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology*, 56, 754-761.
- Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues of for field settings*. Chicago: Rand McNally.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78, 94-104.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley & Sons, Inc.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.

- Feldt, L. S., & Brennan, R. L. (1993). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). Phoenix, AZ: American Council on Education and The Oryx Press.
- Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, 7, 286-299.
- Foster, S. L., & Cone, J. D. (1995). Validity issues in clinical assessment. *Psychological Assessment*, 7, 248-260.
- Gordon, J., Ludlum, J., & Hoey, J. J. (2008). Validating NSSE against student outcomes: Are they related? *Research in Higher Education*, 49, 19-39.
- Harrington, D. (2009). *Confirmatory factor analysis*. New York: Oxford University Press, Inc.
- Hayes, S. C., Nelson, R. O., & Jarrett, R. B. (1987). The treatment utility of assessment: A functional approach to evaluating assessment quality. *American Psychologist*, 42, 963-974.
- Hoyle, R. H., & Smith, G. T. (1994). Formulating clinical research hypotheses as structural equation models: A conceptual overview. *Journal of Consulting and Clinical Psychology*, 62, 429-440.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1-55.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527-535.
- Kinzie, J., & Pennipede, B. S. (2009). Converting engagement results into action. *New Directions for Institutional Research*, 141, 83-96.

- Kline, R.B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: The Guilford Press.
- Koeske, G. F. (1994). Some recommendations for improving measurement validation in social work research. *Journal of Social Service Research*, 18, 43-72.
- Kuh, G. D. (2001a). Assessing what really matters to student learning: Inside the National Survey of Student Engagement. *Change*, 33, 10-17, 66.
- Kuh, G.D. (2001b). *The National Survey of Student Engagement : Conceptual framework and overview of psychometric properties*. Bloomington, IN: Center for Postsecondary Research and Planning.
- Kuh, G. D. (2005). Putting student engagement results to use: Lessons from the field. *Assessment Update*, 17, 12-13.
- Kuh, G.D. (2009). The National Survey of Student Engagement: Conceptual and empirical foundations. *New Directions for Institutional Research*, 141, 5-20.
- Kuh, G. D., Gonyea, R. M., & Palmer, M. (2001). The disengaged commuter student: Fact or fiction? *Commuter Perspectives*, 27, 2-5.
- Kuh, G. D., Hayek, J. C., Carini, R.M., Ouimet, J. A., Gonyea, R. M., & Kennedy, J. (2001). *NSSE technical and norms report*. Bloomington, IN: Indiana University Center for Postsecondary Research and Planning.
- Kuh, G.D., Vesper, N., Connolly, M.R., & Pace, C.R. (1997). *College Student Experiences Questionnaire: Revised norms for the third edition*. Bloomington, IN: Indiana University Center for Postsecondary Research and Planning.

- LaNasa, S. M., Olson, E., & Alleman, N. (2007). The impact of on-campus student growth on first-year student engagement and success. *Research in Higher Education*, 48, 941-966.
- LaNasa, S. M., Cabrera, A. F., & Trangsrud, H. (2009). The construct validity of student engagement: A confirmatory factor analysis approach. *Research in Higher Education*, 50, 315-332.
- Lanyon, R. I., & Goodstein, L. D. (1997). *Personality assessment* (3rd ed.). New York: John Wiley & Sons, Inc.
- Lounsbury, J. W., Gibson, L. W., & Saudargas, R. W. (2006). Scale development. In F. T. L. Leong & J. T. Austin (Eds.), *Psychological Research Handbook: A Guide for Graduate Students and Research Assistants* (pp. 125-146). Thousand Oaks: Sage.
- Mardia, K.V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies, *Sankhya B*, 36, 115-128.
- Messick, S. (1993). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.). Phoenix, AZ: American Council on Education and The Oryx Press.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741-749.
- Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, 38, 171-189.

National Survey of Student Engagement [NSSE]. (2000). *The NSSE 2000 report: National benchmarks of effective educational practice*. Retrieved from

<http://nsse.iub.edu/pdf/NSSE%202000%20National%20Report.pdf>

National Survey of Student Engagement [NSSE]. (2008). *A Pocket Guide to Choosing a College: Are You Asking the Right Questions*. Bloomington: Indiana University Center for Postsecondary Research. Retrieved from

http://nsse.iub.edu/html/pocket_guide_intro.cfm.

National Survey of Student Engagement [NSSE]. (2009a). *Assessment for improvement:*

Tracking student engagement over time—Annual results 2009. Bloomington, IN: Indiana University Center for Postsecondary Research.

National Survey of Student Engagement [NSSE] (2009b). *Teacher education council (TEAC) standards mapped to 2008 NSSE survey questions*. Retrieved from

http://nsse.iub.edu/institute/documents/accred/special2008/TEAC_010509_final.pdf

National Survey of Student Engagement [NSSE] (2010a). *Reliability: 2009 internal consistency*. Retrieved from

http://nsse.iub.edu/pdf/psychometric_portfolio/Reliability_Intercorrelation_2009.pdf

National Survey of Student Engagement [NSSE]. (2010b). *Reliability 2009 student-level test-retest analysis*. Retrieved from

http://nsse.iub.edu/pdf/psychometric_portfolio/2009%20Reliability%20Student-Level%20Test-Retest%20Analysis.pdf

National Survey of Student Engagement [NSSE]. (n.d.(a)). *About NSSE*. Retrieved from

<http://nsse.iub.edu/html/about.cfm>

National Survey of Student Engagement [NSSE]. (n.d.(b)). *NSSE accreditation toolkits*.

Retrieved from <http://nsse.iub.edu/?cid=333>

National Survey of Student Engagement [NSSE]. (n.d.(c)). *NSSE multi-year data analysis guide*.

Retrieved from

http://nsse.iub.edu/2010_Institutional_Report/pdf/NSSE%20MYDAG.pdf

National Survey of Student Engagement [NSSE]. (n.d.(d)). *Survey development*. Retrieved from

http://nsse.iub.edu/html/survey_development.cfm

Pace, C.R. (1984). *Measuring the quality of college student experiences*. Los Angeles:

University of California, Center for the Study of Evaluation.

Pace, C.R. (1990). *The undergraduates: A report of their activities and college experiences in*

the 1980s. Los Angeles: Center for the Study of Evaluation, UCLA Graduate School of Education.

Pascarella, E. T., Seifert, T. A., & Blaich, C. (2008). *Validation of the NSSE benchmarks and deep approaches to learning against liberal arts outcomes*. Paper presented at the Annual Meeting of the Association for the Study of Higher Education.

Pascarella, E. T., Seifert, T. A., & Blaich, C. (2010). How effective are the NSSE benchmarks in predicting important educational outcomes? *Change*, 36, 16-22.

Pascarella, E.T., & Terenzi, P.T. (2005). *How college affects students: A third decade of research* (Vol. 2). San Francisco: Jossey-Bass.

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

- Pike, G. R. (2006b). The dependability of NSSE scalelets for college-and department-level assessment. *Research in Higher Education*, 47, 177-195.
- Porter, S.R. (2009). Do college student surveys have any validity? Paper presented at the meeting of the Association for the Study of Higher Education, Vancouver, Canada.
- Ouimet, J.A., Carini, R.M., Kuh, G.D., and Bunnage, J.C. (2001). *Using focus groups to establish the validity and reliability of a college student survey*. Paper presented at the meeting of Association of Institutional Research, Long Beach, CA.
- Sax, L.J., Astin, A.W., Korn, W.S., & Mahoney, K. (1997). *The American college freshman*. Los Angeles: University of California, Los Angeles.
- Schonrock-Adema, J., Heijne-Penninga, M., Van Hell, E.A., & Cohen-Schotanus, J. (2009). Necessary steps in factor analysis: enhancing validation studies of educational instruments. *Medical Teacher*, 31, e226-e232.
- Smith, G. T., & McCarthy, D. M. (1995). Methodological considerations in the refinement of clinical assessment instruments. *Psychological Assessment*, 7, 300-308.
- Steiger, J. H., & Lind, J. M. (1980). *Statistically based tests for the number of common factors*. Paper presented at the meeting of the Psychometric Society, Iowa City, IA.
- Swerdzewski, P. J., Miller, B. J., & Mitchell, R. L. (2007, October). *Investigating the validity of the National Survey of Student Engagement*. Paper presented at the meeting of the Northeastern Educational Research Association, Rocky Hill, CT.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson.
- Thurstone, L. L. (1947). *Multiple-factor analysis*. Chicago: University of Chicago Press.

- Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika* 38, 1-10.
- Widaman, K.F. (1993). Common factor analysis versus principal component analysis: Differential bias in representing model parameters? *Multivariate Behavioral Research*, 23, 263-311.
- Zhao, C., & Kuh, G.D. (2004). Adding value: Learning communities and student engagement. *Research in Higher Education*, 45, 115-138.

APPENDIX

Table A.1

NSSE Benchmark Items

Item	Description
Level of Academic Challenge (LAC)	
FY: $\alpha = .73$ SR: $\alpha = .76$	
Readagn	Number of assigned textbooks, books, or book-length 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 of fewer than 5 pages
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
Synthesz	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
Workhard	Worked harder than you thought you could to meet an instructor's standards or expectations
Acadpr01	Preparing for class (studying, reading, writing, doing homework or lab work, analyzing data, rehearsing, and other academic activities)
Envschol	Spending significant amounts of time studying and on academic work
Active and Collaborative Learning (ACL)	
FY: $\alpha = .66$ SR: $\alpha = .66$	
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)
†Commproj	Participated in a community-based project (e.g., service learning) as part of a regular course
Oocideas	Discussed ideas from your readings or classes with others outside of class (students, family members, co-workers, etc.)

Table A.1 (continued)

Item	Description
Student-Faculty Interaction (SFI)	
FY: $\alpha = .71$ SR: $\alpha = .74$	
Facgrade	Discussed grades or assignments with an instructor
Facideas	Discussed ideas from your readings or classes with faculty members outside of class
Facplans	Talked about career plans with a faculty member or advisor
Facfeed	Received prompt written or oral feedback from faculty on your academic performance
Facother	Worked with faculty members on activities other than coursework
*Resrch04	Work on a research project with a faculty member outside of course or program requirements
Enriching Educational Experiences (EEE)	
FY: $\alpha = .59$ SR: $\alpha = .66$	
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 of your own
Endivrs	Encouraging contact among students from different economic, social, and racial or ethnic backgrounds
†Cocurr01	Participating in co-curricular activities (organizations, campus publications, student government, fraternity or sorority, intercollegiate or intramural sports, etc.)
Itacadem	Used an electronic medium (listserv, chat group, Internet, instant messaging, etc.) to discuss or complete an assignment
*Intern04	Practicum, internship, field experience, co-op experience, or clinical assignment
*Volntr04	Community service or volunteer work
*Lrncom04	Participate in a learning community or some other formal program where groups of students take two or more classes together
*Forlng04	Foreign language coursework
*Stdabr04	Study abroad
*Indstd04	Independent study or self-designated major

Table A.1 (continued)

Item	Description
*Snrx04	Culminating senior experiences (capstone course, senior project or thesis, comprehensive exam, etc.)
Supportive Campus Environment (SCE)	
FY: $\alpha = .79$ SR: $\alpha = .80$	
Envsocal	Providing the support you need to thrive socially
Envsuprt	Providing the support you need to help you succeed academically
Envnacad	Helping you cope with your non-academic responsibilities (work, family, etc.)
Envstu	Relationships with other students
Envfac	Relationships with faculty members
Envadm	Relationships with administrative personnel and offices

Note. Adapted from NSSE, 2010(a). Sample sizes ranged from 139,815 (FY, SCE) to 175,349 (SY ACL). Items beginning with the prefix “env” generally ask students to indicate the extent to which the institution encourages a particular type of behavior.

An asterisk (*) indicates that an item was not included in the current study because of its nominal scale of measurement.

A dagger (†) indicates that an item was not used in the current study because of non-normality.

Table A.2

Pike's Scalelet Items

Item	Description
Course Challenge	
Workhard	Worked harder than you thought you could to meet an instructor's standards or expectations
Clunprep ^b	Come to class without completing readings or assignments
Exams	Your examinations during the current school year challenged you to do your best work
Acadpr01	Preparing for class (studying, reading, writing, doing homework or lab work, analyzing data, rehearsing, and other academic activities)
Envschol	Spending significant amounts of time studying and on academic work
Writing	
Rewropap	Prepared two or more drafts of a paper or assignment before turning it in
Integrat	Worked on a paper or project that required integrating ideas or information from various sources
†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 of fewer than 5 pages
Higher Order Thinking Skills	
Memorize ^b	Memorizing facts, ideas, or methods from your courses and readings so you can repeat them in pretty much the same form
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
Synthesz	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

Table A.2 (continued)

Item	Description
<i>Active Learning</i>	
Clquest	Asked questions in class or contributed to class discussions
Clpresen	Made a class presentation
†Commproj	Participated in a community-based project as part of a regular course
<i>Collaborative Learning</i>	
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)
Oocideas	Discussed ideas from your readings or classes with others outside of class (students, family members, coworkers, etc.)
<i>Course Interaction</i>	
Facgrade	Discussed grades or assignments with an instructor
Facideas	Discussed ideas from your readings or classes with faculty members outside of class
Facfeed	Received prompt feedback from faculty on your academic performance (written or oral)
<i>Out-of-Class Interaction</i>	
Facplans	Talked about career plans with a faculty member or advisor
Facother	Worked with faculty members on activities other than coursework (committees, orientation, student-life activities, etc.)
*Resrch04	Work on a research project with a faculty member outside of course or program requirements
<i>Varied Experiences</i>	
*Intern04	Practicum, internship, field experience, co-op experience, or clinical assignment
*Volntr04	Community service or volunteer work

Table A.2 (continued)

Item	Description
*Lrncom04	Participate in a learning community or some other formal program where groups of students take two or more classes together
*Forlng04	Foreign language coursework
*Stdabr04	Study abroad
*Indstd04	Independent study or self-designated major
*Snrx04	Culminating senior experiences (capstone course, senior project or thesis, comprehensive exam, etc.)
†Cocurr01	Participating in co-curricular activities (organizations, campus publications, student government, social fraternity or sorority, intercollegiate or intramural sports, etc.)
Envevent ^a	Attending campus events and activities (special speakers, cultural performances, athletic events, etc.)
Information Technology	
Itacadem	Used an electronic medium (list-serv, chat group, Internet, etc.) to discuss or complete an assignment
Email	Used e-mail to communicate with an instructor
Envcompt	Using computers in academic work
Diversity	
Divrstud	Had serious conversations with students of a different race or ethnicity than your own
Diffstu2	Had serious conversations with students who differ from you in terms of their religious beliefs, political opinions, or personal values
Envdivrs	Encouraging contact among students from different economic, social, and racial or ethnic backgrounds
Support for Student Success	
Envsuprt	Providing the support you need to help you succeed academically
Envnacad	Helping you cope with your non-academic responsibilities (work, family, etc.)
Envsocial	Providing the support you need to thrive socially

Table A.2 (continued)

Item	Description
Interpersonal Environments	
Envstu	Relationships with other students
Envfac	Relationships with faculty members
Envadm	Relationships with administrative personnel and offices

Note. Table adapted from Pike (2006b). Scalelets in italics were not included in analyses because they consisted of fewer than three items.

^a Items beginning with the prefix “env” ask students to indicate the extent to which the institution encourages a particular type of behavior, with the exception of items included in the *Interpersonal Environments* scalelet.

^b Item is reverse-scored.

An asterisk (*) indicates that an item was not included in the current study because of its nominal scale of measurement.

A dagger (†) indicates that an item was not used in the current study because of non-normality.

Table A.3

Normality Assessment

Variable	Skew	c.r.	Kurtosis	c.r.
Envdivrs	.096	1.70	-.67	-6.03
Writesml	.484	8.59	-.39	-3.49
Envschol	-.47	-8.50	-.06	-.55
Acadpr01	.74	13.23	.15	1.33
Oocideas	.00	.07	-.78	-7.00
Tutor	1.01	18.09	.47	4.24
Occgrp	.05	1.01	-.80	-7.17
Classgrp	.17	3.16	-.56	-5.03
Clpresen	.43	7.63	-.20	-1.79
Clquest	.03	.69	-.98	-8.75
Envsocal	.30	5.32	-.53	-4.70
Envnacad	.77	13.84	.06	.60
Envsuprt	-.24	-4.40	-.33	-2.93
Envadm	-.18	-3.26	-.44	-3.92
Envfac	-.51	-9.18	.12	1.11
Envstu	-.87	-15.52	.67	5.99
Writemid	.90	16.03	1.58	14.03
Readasgn	.22	3.89	-.49	-4.41
Applying	-.58	-10.33	-.40	-3.60
Evaluate	-.25	-4.55	-.58	-5.20
Synthesz	-.25	-4.53	-.63	-5.59
Analyze	-.55	-9.86	-.28	-2.48
Workhard	.06	1.23	-.49	-4.36
Facother	1.22	21.78	.73	6.51
Facfeed	.05	.94	-.41	-3.71
Facideas	.85	15.15	.22	1.96
Facplans	.37	6.62	-.50	-4.46
Facgrade	.07	1.30	-.83	-7.37
Diffstu2	-.11	-2.10	-.94	-8.42
Divrstud	.08	1.53	-.97	-8.60
Itacadem	-.07	-1.30	-1.02	-9.09
Exams	-.95	-17.02	1.80	16.14
Rewropap	.17	3.15	-.97	-8.69
Integrat	-.45	-8.10	-.58	-5.27
Envevent	-.23	-4.11	-.56	-5.03

Table A.3 (continued)

Variable	Skew	c.r.	Kurtosis	c.r.
Email	-.68	-12.32	-.71	-6.39
Envcompt	-1.03	-18.60	.43	3.85
Memorize_rev	.38	6.87	-.51	-4.60
Clunprep_rev	-.63	-11.37	.36	3.27

Table A.4

Benchmarks: Internal Consistency

(a)

Level of Academic Challenge (LAC) $\alpha = .71$

	Readasgn	Writemid	Writesml	Analyze	Synthesz	Evaluate	Applying	Workhard	Acadpr01	Envschol	Corrected Item-Total <i>r</i>
Readasgn	*	.31	.27	.11	.11	.08	.03	.10	.22	.09	.27
Writemid		*	.34	.21	.21	.17	.13	.17	.15	.05	.35
Writesml			*	.13	.12	.13	.10	.06	.11	.03	.25
Analyze				*	.61	.49	.49	.27	.20	.21	.53
Synthesz					*	.57	.50	.30	.22	.23	.56
Evaluate						*	.47	.30	.13	.19	.48
Applying							*	.24	.20	.22	.46
Workhard								*	.26	.26	.39
Acadpr01									*	.23	.33
Envschol										*	.30

Table A.4 (continued)

(b)

Active and Collaborative Learning (ACL) $\alpha = .62$

	Clquest	Clpresen	Classgrp	Occgrp	Tutor	Oocideas	Corrected Item-Total <i>r</i>
Clquest	*	.27	.11	.14	.22	.27	.33
Clpresen		*	.24	.40	.15	.22	.43
Classgrp			*	.29	.12	.15	.29
Occgrp				*	.24	.22	.43
Tutor					*	.18	.29
Oocideas						*	.34

(c)

Student-Faculty Interaction (SFI) $\alpha = .75$

	Facgrade	Facideas	Facplans	Facfeed	Facother	Corrected Item- Total <i>r</i>
Facgrade	*	.41	.46	.33	.29	.51
Facideas		*	.49	.36	.42	.59
Facplans			*	.32	.39	.58
Facfeed				*	.22	.41
Facother					*	.45

Table A.4 (continued)

(d)

Enriching Educational Experience (EEE) $\alpha = .57$

	Diffstu2	Divrstud	Envdivrs	Itacadem	Corrected Item-Total <i>r</i>
Diffstu2	*	.68	.15	.20	.53
Divrstud		*	.16	.19	.52
Envdivrs			*	.09	.17
Itacadem				*	.22

(e)

Supportive Campus Environment (SCE) $\alpha = .73$

	Envsocal	Envsuprt	Envnacad	Envstu	Envfac	Envadm	Corrected Item- Total <i>r</i>
Envsocal	*	.40	.60	.31	.25	.25	.49
Envsuprt		*	.42	.25	.38	.33	.51
Envnacad			*	.20	.24	.26	.46
Envstu				*	.35	.29	.41
Envfac					*	.53	.55
Envadm						*	.51

Table A.5

Benchmarks: Analysis of Residuals

Benchmark Item	Number of Standardized Residual Covariances > 2
Itacadem	12
Divrstud	4
Diffstu2	5
Facgrade	6
Facplans	6
Facideas	10
Facfeed	14
Facother	15
Workhard	16
Analyze	11
Synthesz	5
Evaluate	4
Applying	11
Readasgn	10
Writemid	13
Envstu	15
Envfac	11
Envadm	11
Envsuprt	10
Envnacad	11
Envsocial	7
Clquest	11
Clpresen	11
Classgrp	10
Occgrp	12
Tutor	9
Oocideas	10
Acadpr01	9
Envschol	10
Writesml	10
Envdivrs	20

Table A.6

Benchmarks: Modification Indices

Freely estimating these...			...Would result in a χ^2 change of approximately:
Clpresen_res	&	Occgrp_res	110.5
Envsocal_res	&	Envdivrs_res	109.6
Envnacad_res	&	Envdivrs_res	117.4
Envnacad_res	&	Envsocal_res	282.4
Envsuprt_res	&	Envdivrs_res	117.1
Envsuprt_res	&	Envschol_res	161.7
Envfac_res	&	Envsocal_res	105.7
Envfac_res	&	Envnacad_res	109.6
Envfac_res	&	Envadm_res	218.6
Writemid_res	&	Writesml_res	172.4
Readasgn_res	&	Writesml_res	114.9
Readasgn_res	&	Writemid_res	149.7
Envdivrs	&	SCE	387.1

Table A.7

Benchmarks:Parameters

Benchmark Item	Estimate	Communality
EEE		
Divrstud	0.81	0.66
Diffstu2	0.83	0.69
Itacadem	0.28	0.08
Envdivrs	0.21	0.04
SFI		
Facgrade	0.60	0.36
Facplans	0.68	0.46
Facideas	0.69	0.48
Facfeed	0.52	0.27
Facother	0.56	0.31
LAC		
Workhard	0.45	0.20
Analyze	0.73	0.53
Synthesz	0.77	0.59
Evaluate	0.69	0.48
Applying	0.65	0.42
Readasgn	0.18	0.03
SCE		
Envstu	0.46	0.21
Envfac	0.60	0.36
Envadm	0.55	0.30
Envsuprt	0.62	0.38
Envnacad	0.62	0.38
Envsocal	0.64	0.41
ACL		
Clquest	0.51	0.26

Table A.7 (continued)

Benchmark Item	Estimate	Communality
Clpresen	0.47	0.22
Classgrp	0.33	0.11
Occgrp	0.46	0.21
Tutor	0.41	0.17
Oocideas	0.56	0.31
	LAC	
Writesml	0.20	0.04
Writemid	0.30	0.09
Acadpr01	0.31	0.10
Envschol	0.32	0.10

Table A.8

Scalelets: Internal Consistency

(a)

Course Challenge $\alpha = .57$

						Corrected Item-Total <i>r</i>
	Workhard	Clunprep_rev	Exams	Acadpr01	Envschol	
Workhard	*	.10	.31	.27	.28	.38
Clunprep_rev		*	.13	.22	.09	.22
Exams			*	.26	.34	.39
Acadpr01				*	.24	.38
Envschol					*	.37

(b)

Writing $\alpha = .54$

					Corrected Item- Total <i>r</i>
	Rewropap	Integrat	Writemid	Writesml	
Rewropap	*	.38	.18	.09	.29
Integrat		*	.30	.17	.41
Writemid			*	.33	.39
Writesml				*	.26

(c)

Higher Order Thinking $\alpha = .68$

						Corrected Item- Total <i>r</i>
	Memorize_rev	Analyze	Synthesz	Evaluate	Applying	
Memorize_rev	*	-.07	-.02	-.02	-.01	-.04
Analyze		*	.62	.49	.49	.58
Synthesz			*	.57	.50	.64
Evaluate				*	.47	.57
Applying					*	.54

Table A.8 (continued)

(d)

Diversity $\alpha = .61$

	Divrstud	Diffstu2	Envdivrs	Corrected Item- Total <i>r</i>
Divrstud	*	.68	.17	.57
Diffstu2		*	.15	.56
Envdivrs			*	.17

(e)

Collaborative Learning $\alpha = .68$

	Classgrp	Occgrp	Tutor	Oocideas	Facgrade	Facideas	Facfeed	Corrected Item- Total <i>r</i>
Classgrp	*	.29	.11	.15	.17	.15	.15	.27
Occgrp		*	.23	.23	.23	.23	.15	.38
Tutor			*	.18	.20	.31	.18	.33
Oocideas				*	.26	.35	.29	.40
Facgrade					*	.40	.34	.45
Facideas						*	.36	.51
Facfeed							*	.40

(f)

Course Interaction $\alpha = .63$

	Facgrade	Facideas	Facfeed	Corrected Item- Total <i>r</i>
Facgrade	*	.40	.34	.45
Facideas		*	.36	.47
Facfeed			*	.42

Table A.8 (continued)

(g)

Instructional Technology $\alpha = .49$

				Corrected Item-Total <i>r</i>
	Itacadem	Email	Envcompt	
Itacadem	*	.30	.22	.34
Email		*	.22	.34
Envcompt			*	.27

(h)

Support for Student Services $\alpha = .73$

				Corrected Item-Total <i>r</i>
	Envsuprt	Envnacad	Envsocal	
Envsuprt	*	.42	.41	.46
Envnacad		*	.60	.61
Envsocal			*	.60

(i)

Interpersonal Environment $\alpha = .66$

				Corrected Item- Total <i>r</i>
	Envstu	Envfac	Envadm	
Envstu	*	.35	.29	.36
Envfac		*	.53	.56
Envadm			*	.50

Table A.9

Scalelets: Analysis of Residuals

Scalelet Item	Number of Standardized Residual Covariances with z-scores > 2
Memorize_rev	12
Envdivrs	23
Envadm	10
Envfac	7
Envstu	16
Diffstu2	6
Divrstud	4
Envsocal	8
Envnacad	11
Envsuprt	11
Envcompt	14
Email	12
Itacadem	4
Facfeed	6
Facideas	7
Facgrade	9
Oocideas	3
Tutor	7
Occgrp	11
Classgrp	9
Applying	8
Evaluate	7
Synthesz	3
Analyze	7
Writesml	7
Writemid	6
Integrat	5
Rewropap	6
Envschol	9
Acadpr01	8
Exams	11
Clunprep_rev	9
Workhard	12

Table A.10

Scalelets: Modification Indices

Freely estimating these...			...Would result in a χ^2 change of approximately:
Facgrade_res	&	Email_res	201.93
Writemid_res	&	Writesml_res	121.15
Envschol_res	&	Envsuprt_res	129.53
Envdivrs	&	Support for Student Services	498.46
Envdivrs	&	Course Challenge	120.06
Envdivrs	&	Interpersonal Environment	161.32

Table A.11

Scalelets: Parameters

Scalelet Item	Estimate	Communality
Course Challenge		
Workhard	0.64	0.40
Clunprep_rev	0.21	0.04
Exams	0.52	0.27
Acadpr01	0.45	0.20
Envschol	0.50	0.25
Writing		
Rewropap	0.51	0.26
Integrat	0.71	0.50
Writemid	0.44	0.19
Writesml	0.29	0.08
Higher-Order Thinking Skills		
Analyze	0.75	0.56
Synthesz	0.79	0.62
Evaluate	0.70	0.49
Memorize_rev	-0.07	0.00
Applying	0.66	0.44
Collaborative Learning		
Classgrp	0.31	0.10
Occgrp	0.43	0.18
Tutor	0.37	0.14
Oocideas	0.60	0.36
Course Interaction		
Facgrade	0.63	0.40

Table A.11 (continued)

Scalelet Item	Estimate	Communality
Facideas	0.61	0.37
Facfeed	0.57	0.32
Information Technology		
Itacadem	0.49	0.24
Email	0.61	0.37
Envcompt	0.40	0.16
Support for Student Success		
Envsuprt	0.60	0.36
Envnacad	0.75	0.56
Envsocal	0.74	0.55
Diversity		
Divrstud	0.80	0.64
Diffstu2	0.84	0.71
Envdivrs	0.21	0.04
Interpersonal Environment		
Envstu	0.48	0.23
Envfac	0.77	0.59
Envadm	0.67	0.45

Table A.12

Factor Analysis: Initial Pattern Matrix

	<u>Factor</u>									
	1	2	3	4	5	6	7	8	9	10
Divrstud	-0.04	0.06	0.06	0.78	-0.04	-0.02	-0.03	0.02	-0.02	0.00
Diffstu2	0.02	-0.02	0.00	0.90	-0.03	-0.01	-0.03	-0.01	-0.03	-0.01
Facgrade	-0.04	0.25	0.02	0.00	-0.07	0.63	0.01	-0.06	-0.06	0.00
Facplans	-0.02	0.49	0.01	-0.02	0.10	0.27	0.05	-0.01	-0.06	-0.07
Facideas	0.05	0.67	0.01	0.01	0.04	0.07	-0.01	-0.06	-0.01	0.00
Facfeed	0.11	0.16	0.10	0.02	0.20	0.16	-0.06	-0.10	0.06	0.10
Facother	0.01	0.63	0.10	-0.04	0.01	-0.05	-0.06	0.15	0.01	-0.10
Workhard	0.04	0.19	0.01	-0.01	-0.07	0.13	0.32	0.06	-0.04	0.17
Analyze	0.78	-0.03	-0.06	0.01	0.00	-0.02	0.07	-0.01	0.04	-0.08
Synthesz	0.86	0.06	0.00	-0.02	-0.06	-0.07	0.02	-0.07	-0.01	-0.01
Evaluate	0.70	0.01	0.06	0.01	-0.04	-0.02	-0.09	-0.01	-0.05	0.10
Applying	0.64	0.04	0.00	-0.01	0.03	-0.09	0.08	0.11	-0.04	-0.06
Readasgn	-0.10	-0.07	0.02	0.04	0.03	0.02	0.14	-0.14	0.53	0.05
Writemid	0.05	0.06	0.01	-0.06	-0.05	0.01	-0.02	-0.01	0.55	0.07
Envstu	-0.10	-0.04	0.09	0.06	0.30	-0.03	0.12	0.33	0.07	-0.09
Envfac	0.01	0.03	-0.07	-0.01	0.95	-0.04	-0.11	-0.04	0.02	-0.02
Envadm	-0.06	0.08	0.06	-0.06	0.62	-0.13	0.08	0.06	-0.10	0.02
Envsuprt	0.05	-0.14	0.45	0.01	0.24	0.07	0.17	-0.08	-0.04	0.00
Envnacad	-0.04	0.15	0.77	-0.05	-0.06	-0.05	-0.04	-0.01	-0.01	0.06
Envsocial	-0.01	0.06	0.81	0.00	-0.03	-0.08	-0.12	0.04	0.09	0.02
Clquest	0.03	0.29	-0.07	0.08	0.09	0.08	0.01	-0.02	0.14	0.04
Clpresen	0.03	0.14	-0.05	-0.08	0.01	0.05	-0.10	0.45	0.09	0.14
Classgrp	0.05	0.02	0.07	0.04	-0.01	0.01	-0.05	0.33	-0.06	0.07
Occgrp	-0.02	0.11	-0.08	0.01	-0.01	0.01	0.07	0.64	-0.08	0.04
Tutor	-0.01	0.44	0.02	0.08	-0.05	-0.14	0.06	0.17	0.00	-0.04
Oocideas	0.10	0.10	-0.09	0.31	0.10	0.07	0.05	0.03	0.06	0.02
Acadpr01	0.00	0.20	-0.09	0.02	-0.05	-0.20	0.53	-0.04	0.14	0.09
Envschol	0.08	-0.12	0.10	-0.05	-0.05	0.16	0.48	0.04	-0.01	-0.04
Writesml	0.01	0.02	0.04	-0.03	-0.05	0.01	-0.02	0.05	0.62	-0.14
Envdivrs	0.04	-0.04	0.60	0.09	-0.03	0.03	0.06	-0.07	-0.02	0.01
Clunprep_rev	-0.00	-0.01	-0.03	-0.01	0.06	-0.08	0.25	-0.18	0.03	0.21
Exams	0.03	-0.08	-0.01	-0.02	0.03	0.07	0.53	0.05	-0.04	0.02
Rewropap	-0.07	-0.02	0.07	-0.02	-0.03	-0.05	0.12	0.04	-0.07	0.66
Integrat	0.06	-0.17	0.00	0.04	0.03	0.06	-0.04	0.20	0.06	0.57
Itacadem	0.16	-0.00	-0.05	0.03	0.01	0.23	-0.01	0.13	0.00	0.00
Email	-0.07	-0.02	-0.07	-0.02	-0.05	0.79	-0.04	0.02	0.03	0.02

Table A.12 (continued)

	<u>Factor</u>									
	1	2	3	4	5	6	7	8	9	10
Envcompt	0.20	-0.16	0.13	-0.00	0.04	0.23	-0.00	0.11	0.10	-0.10
Memorize_rev	0.11	-0.02	-0.04	-0.02	0.11	-0.23	-0.14	-0.05	-0.04	0.12

Table A.13

Final Factor Analysis Solution

	<u>Factor</u>					
	Communication	Environmental Support	Critical Thinking	Openness to Diversity	Academic Scholarship	Literacy
Email	.49					
Facgrade	.57					
Facideas	.60					
Facother	.63					
Workhard	.38					
Clpresen	.50					
Occgrp	.48					
Tutor	.39					
Envfac		.38				
Envadm		.41				
Envsuprt		.61				
Envnacad		.72				
Envsocal		.75				
Envdivrs		.59				
Analyze			.73			
Synthesz			.80			
Evaluate			.66			
Applying			.61			
Divrstud				.76		
Diffstu2				.89		
Oocideas				.33		
Acadpr01					.34	
Envschol					.55	
Exams					.57	
Readasgn						.57
Writemid						.58
Writesml						.53

Table A.14

Six-Factor Model: Analysis of Residuals

Scalelet Item	Number of Standardized Residual Covariances with z-scores > 2
Diffstu2	2
Oocideas	21
Divrstud	4
Applying	9
Occgrp	7
Envdivrs	8
Envsocal	10
Envnacad	13
Envsuprt	5
Envadm	12
Envfac	19
Exams	10
Envschol	10
Acadpr01	8
Tutor	11
Clpresen	12
Workhard	7
Facother	13
Facideas	10
Facgrade	10
Email	10
Writesml	6
Writemid	4
Readasgn	10
Evaluate	6
Synthesize	6
Analyze	5

Table A.15

Six-Factor Model: Modification Indices

Freely estimating these...			...Would result in a χ^2 change of approximately:
Envnacad_res	&	Envsocial_res	158.98
Envfac_res	&	Envadm_res	420.43
Envschol_res	&	Envsuprt_res	104.86
Clpresen_res	&	Occgrp_res	112.72
Email_res	&	Facgrade_res	233.79
Workhard	&	Academic Scholarship	126.36

Table A.16

Six-Factor Model: Parameters

Scalelet Item	Estimate	Communality
Communication		
Email	0.38	0.14
Facgrade	0.53	0.28
Facideas	0.51	0.26
Facother	0.46	0.21
Workhard	0.42	0.18
Clpresen	0.38	0.14
Occgrp	0.41	0.17
Tutor	0.32	0.10
Environmental Support		
Envfac	0.59	0.35
Envadm	0.68	0.46
Envsuprt	0.54	0.29
Envnacad	0.63	0.40
Envsocal	0.61	0.37
Envdivrs	0.62	0.38
Critical Thinking		
Analyze	0.52	0.27
Synthesz	0.65	0.42
Evaluate	0.57	0.32
Applying	0.52	0.27
Openness to Diversity		
Divrstud	0.74	0.55
Diffstu2	0.81	0.66
Oocideas	0.36	0.13

Table A.16 (continued)

Scalelet Item	Estimate	Communality
Academic Scholarship		
Acadpr01	0.70	0.49
Envschol	0.41	0.17
Exams	0.58	0.34
Literacy		
Readasgn	0.47	0.22
Writemid	0.56	0.31
Writesml	0.58	0.34

VITA

Shelley Esquivel was born in Galveston, Texas and raised in Ashland, Virginia. In 1998, she graduated from Patrick Henry High School and enrolled as a student at Florida State University (FSU). After completing three semesters at FSU, Shelley transferred to the Psychology program at the University of Nebraska, Omaha (UNO). Upon completion of her B.A. in Psychology in 2004, Shelley enrolled in a master's program at UNO, where she received her M.A. in Developmental Psychology in 2008. While a master's student, Shelley had the opportunity to serve as a teaching assistant for several courses, including Psychology, Advanced Psychology, and Statistics. After obtaining her M.A. in Psychology, Shelley enrolled in the Educational Psychology and Research doctoral program at The University of Tennessee, Knoxville (UT). While a doctoral student, Shelley was a graduate assistant at the Office of Institutional Research and Assessment (OIRA). Shelley's interests include institutional research, assessment, instrument validation, statistics, and survey research methodology.