



Abstract

This study utilized generalizability theory to assess the context where the National Survey of Student Engagement's (NSSE) summary measures, the Engagement Indicators, produce dependable group-level means. The dependability of NSSE group means is an important topic for the higher education assessment community given its wide utilization and usage in institutional assessment and accreditation. We found that the Engagement Indicators produced dependable group means for an institution derived from samples as small as 25 to 50 students. Furthermore, we discuss how the assessment community should use NSSE data.

AUTHORS

Kevin Fosnacht, Ph.D.
Indiana University,
Bloomington

Robert M. Gonyea, Ed.D.
Indiana University,
Bloomington

The Dependability of the Updated NSSE: A Generalizability Study

Despite decades of dialogue, higher education still struggles with assessing the quality of undergraduate education and no longer enjoys respectful deference from governments, media, and the public who are collectively anxious about cost and quality. Such anxieties have stimulated considerable pressure for assessment and accountability. The dominant paradigm focusing on resources and reputation – most visible in the *U.S. News and World Report* rankings – has been roundly criticized for its neglect of students' educational experiences (Carey, 2006; McGuire, 1995). In response, higher education leaders, researchers, and assessment professionals have explored many ways for higher education to improve – through reforming the curriculum, faculty development, and improved assessment (Association of American Colleges and Universities, n.d.; Barr & Tagg, 1995; Gaston, 2010; Lumina Foundation, 2011). In recent years, the measurement of student engagement has emerged as a viable alternative for institutional assessment, accountability, and improvement efforts. Student engagement represents collegiate quality in two critical ways. The first is the amount of time and effort students put into their coursework and other learning activities, and the second is how the institution allocates resources, develops the curriculum, and promotes enriching educational activities that decades of research studies show promote student learning. (Kuh, 2003, 2009; Kuh, Hayek, Carini, Ouimet, Gonyea, & Kennedy, 2001; McCormick, Kinzie, & Gonyea, 2013).

CORRESPONDENCE

Email
kfosnach@indiana.edu

The National Survey of Student Engagement (NSSE) collects information at hundreds of bachelor's-granting universities to estimate how students spend their time and how their educational experiences are shaped. Institutions use NSSE primarily in two ways. The first is to compare, or benchmark, their students' responses with those of students at other institutions. Such an approach provides the institution with diagnostic information about how their students are learning, and which aspects of the undergraduate experience have been effective and which are in need of improvement. The second way institutions use NSSE is to assess subgroups of their students to determine how student

engagement varies *within* the institution and to uncover areas for institutional improvement for groups such as first-generation students, part-time students, adult and commuter students, students enrolled in different majors, transfer students, and so on. Both of these approaches utilize NSSE scores by comparing the aggregate score of one group with that of another group, whether they be different institutions or different types of students within the same institution.

Thus, the NSSE instrument depends foremost on its reliability at the group level, and upon its ability to generalize an outcome to the aggregated group. The reliability of a group mean score requires methodological techniques that can account for and identify multiple sources of error (Pike, 1994, 2013). Consequently, this paper explores the notion that generalizability theory (GT) may provide the proper methodological framework to assess the dependability of benchmarking instruments such as NSSE, and uses GT to investigate the number of students needed to produce a reliable group mean for the NSSE Engagement Indicators. Finally, the appropriate uses of NSSE data are discussed in light of the study's findings.

In recent years, the measurement of student engagement has emerged as a viable alternative for institutional assessment, accountability, and improvement efforts.

Updating NSSE

Since NSSE's initial launch in 2000, higher education scholars have learned more about collegiate activities and practices that positively influence student outcomes (Kuh, Kinzie, Cruce, Shoup & Gonyea, 2006; McClenney & Marti, 2006; Pascarella, Seifert, & Blaich, 2010; Pascarella & Terenzini, 2005). Many areas of higher education are seeing growth, innovation, and rapid adoption of new ideas such as distance learning and other technological advances. To meet these challenges and improve the utility and actionability of its instrument, NSSE introduced an updated version in 2013, which both refines its existing measures and incorporates new measures related to emerging practices in higher education (NSSE, 2018b). The new content includes items investigating quantitative reasoning, interactions among diverse populations, learning strategies, and teaching practices. Additionally, the update provides the opportunity to improve the clarity and consistency of the survey's language and to improve the properties of the measures derived from the survey. Despite these changes, the updated instrument is consistent with the purpose and design of the original version of NSSE (Kuh et al., 2001), as it continues to focus on whether institutions emphasize participation in effective educational practices, and is administered to samples of first-year students and seniors at various types of baccalaureate-granting institutions.

Validity of NSSE

With the updated survey, NSSE continues its core purpose of providing institutions with valid and reliable assessment information for the improvement of the educational experience such as helping faculty and senior academic leaders to shape faculty development programs, revise curricula, or develop student support programs. Studies that link student engagement to university outcomes such as critical thinking, moral development, and leadership capacity, or to other indicators of success such as grades, persistence, and graduation, give credence to NSSE's validity and support such valid uses of the data.

For example, research has found positive associations with persistence (Hughes & Pace, 2003; Kuh, 2008; Kuh et al., 2008; McClenney & Marti, 2006), critical thinking (Loes, Pascarella, & Umbach, 2012), GRE scores (Carini, Kuh, & Klein, 2006), moral reasoning (Mayhew et al., 2012), and need for cognition (Padgett et al., 2010). Using institution-level data, NSSE benchmarks had at least one significant positive association with institution-level outcome scores (effective reasoning and problem-solving, moral character, inclination to inquire and lifelong learning, intercultural effectiveness, and personal well-being) for first-year students after controlling for pre-test outcome scores (Pascarella, Seifert, & Blaich, 2010).

Prior research has supported the use of self-reported data on university students (see Pace, 1985 and Pike, 2011), although some (e.g., Porter, 2011) have raised questions about the validity of university student surveys. Cited concerns included a lack of a sufficient

theoretical basis for survey content, difficulties in the response process, the lack of a factor structure and adequate reliability for NSSE's benchmarks, and poor relationships between measures of student engagement and direct observations of the same behavior. In response, NSSE researchers explain that while the student engagement survey items are supported in the literature, the survey was created for institutional assessment, not for theory building or testing of a narrow theoretical construct. Also, students' ability to respond to the survey items has been established by extensive testing with hundreds of students at dozens of institutions using focus groups and cognitive interviews. For a more comprehensive discussion of NSSE's validity, see McCormick and colleagues (2013) and NSSE's (2018c) psychometric portfolio.

Generalizability Theory

Generalizability Theory, first detailed in a monograph by Cronbach, Gleser, Nanda, and Rajaratnam (1972), is a conceptual framework useful in determining the reliability and dependability of measurements. Unlike reliability coefficients such as Cronbach's α that provide a single statistic, GT provides a framework for determining the situations where drawing inferences from their samples would be appropriate. Researchers and assessment professionals can then use this information to design a study, or inferences can be responsibly inferred from their existing data. GT is perhaps best described in relation to classical test theory (CTT) where a person's true score (T) on an item or test is composed of their observed score (X) and measurement error (e): $T=X+e$. Thus, CTT focuses on determining the error of a measurement. In contrast, GT recognizes that multiple sources of error may exist and examines their magnitude rather than focusing on a single overall error score. These potential sources of error (e.g., individuals, raters, items, and occasions) are referred to as facets. More concretely, an error could be due to a student randomly guessing the correct answer on a test, differences in the calibration of a scale, or the implicit biases of a judge or rater. The theory assumes that any observation or data point is drawn from a universe of possible observations. For example, an item on a survey is assumed to be sampled from a universe of comparable items, just as individuals are sampled from a larger population. Consequently, the notion of reliability in CTT is replaced by the question of the "accuracy of generalization or generalizability" to a larger universe (Cronbach et al., 1972, p. 15).

As a methodological theory, GT is intimately associated with its methods. Generalizability Theory utilizes analysis of variance (ANOVA) which analyzes the amount of variation in a measure attributable to groups of people, test items, schools, or other things of interest to a researcher. In the GT context, ANOVA is used to estimate the magnitude of the variance components associated with the types of error identified by the researcher. However, it is important to note that while GT uses ANOVA, it departs from the traditional uses of ANOVA through its focuses on variance components, not testing statistical significance. The researcher subsequently uses the variance components to calculate the generalizability coefficient, which is analogous to the reliability coefficient in CTT. The generalizability coefficient is a type of intraclass correlation coefficient (which measures the proportion of total variance attributable to within-group differences). However, in the generalizability coefficient the true score variance of CTT is replaced with the universe score variance focused on in GT (Kane & Brennan, 1977). GT also distinguishes between a generalizability (G) study and a decision (D) study. The G-study uses ANOVA to estimate the variance components used to calculate the generalizability coefficient. The components can also be used in a D-study to estimate the generalizability coefficient in different contexts. The D-study allows a researcher to efficiently optimize a study or to determine the conditions under which a score is generalizable.

Due to the focus on groups in educational assessment, GT makes important contributions to determining the validity of surveys, such as NSSE. The flexibility of GT allows researchers to determine the conditions under which group means will be accurate and dependable. This is in contrast to the methods based on CTT that look at the internal consistency of a set of items (e.g., Cronbach's α), but fail to identify the conditions under which a measure is accurate. This weakness of CTT approaches may lead well-intentioned researchers to use a measure under conditions where its validity is questionable. Despite the benefits of GT, it has been underutilized in higher education research even after Pike's (1994) work that introduced GT and its methods to the field.

Thus, CTT focuses on determining the error of a measurement. In contrast, GT recognizes that multiple sources of error may exist and examines their magnitude rather than focusing on a single overall error score.

Research Questions

Guided by GT, the study answered the following questions:

1. How dependable are the NSSE Engagement Indicators?
2. How many students are needed to produce a dependable group mean for the NSSE Engagement Indicators?

Methods

Data

The study utilized data from the 2013 NSSE administration. The survey was administered to 334,808 first-year students and seniors at 568 baccalaureate-granting institutions in the United States in the winter and spring of 2013. The characteristics of the institutions and respondents are available from NSSE (2013). The characteristics of the institutions roughly mirror the U.S. landscape, although public institutions and larger master's colleges and universities were overrepresented. Baccalaureate Colleges – Diverse Fields were slightly underrepresented in the dataset. Approximately, two out of three respondents were female, the same proportion was White, and the vast majority enrolled as full-time students. The average institutional response rate was 30%, which prior research using NSSE data has shown to produce estimates that can be generalized to the broader population of students within an institution accurately (Fosnacht, Sarraf, Howe, & Peck, 2017).

The measures used in the study were the survey items that comprised the NSSE Engagement Indicators (EI), groups of related items designed by NSSE researchers to measure the extent to which an institution's environment promotes effective educational practices. The ten indicators are Higher-Order Learning (HO; 4 items), Reflective & Integrative Learning (RI; 7 items), Learning Strategies (LS; 3 items), Quantitative Reasoning (QR; 3 items), Collaborative Learning (CL; 4 items), Discussions with Diverse Others (DD; 4 items), Student-Faculty Interaction (SF; 4 items), Effective Teaching Practices (ET; 5 items), Quality of Interactions (QI; 5 items), and Supportive Environment (SE; 8 items). The full list of the items that comprise the EIs is available from NSSE (2018a), and the abbreviations used match those used by NSSE in its reporting to institutions. The QI items had a "not applicable" option which was recoded to missing for this analysis. All items within each indicator shared the same response set and were not recoded (except for the QI items).

Analyses

Guided by GT, the study examined the group mean generalizability of the NSSE Engagement Indicators at the institution level. We performed the following procedures to assess the generalizability of each EI by class. We identified two facets, students and items, as potential sources of error for the indicators. Additionally, students in our sample were nested within institutions due to the design of NSSE. Thus, the G-study portion of our analyses which estimated the variance components utilized a split-plot, random effects ANOVA design, where students were nested within institutions and crossed with survey items (see Kirk, 2013 for more details on split-plot ANOVA designs). In this design, each institution has a different set of students, but all students answered the same items. The design was also balanced, with 50 students randomly selected from each institution. The value of 50 was selected to maximize the number of students and institutions included in the study after the exclusion of cases with missing data. The mathematical model of the ANOVA was:

$$X_{usi} = \mu + \alpha_u + \pi_{s(u)} + \beta_i + \alpha\beta_{ui} + \beta\pi_{is(u)} + e_{usi}$$

Due to the focus on groups in educational assessment, GT makes important contributions to determining the validity of surveys, such as NSSE. The flexibility of GT allows researchers to determine the conditions under which group means will be accurate and dependable.

Where,

X_{usi} = Response by student s in institution u on item i

μ = grand mean

α_u = effect for institution u

$\pi_{s(u)}$ = effect for student s nested within institution u

β_i = effect for item i

$\alpha_u \beta_i$ = institution by item interaction

$\beta_i \pi_{s(u)}$ = item by student, nested within institution, interaction, and

e_{usi} = error term.

Apart from the grand mean, each of the parameter estimates varies by institution, student, and/or item. This variation allows for the estimation of the variance components which decompose the total model variation into portions attributable to each effect. We used the G1 program for SPSS to analyze the data and estimate the variance components (Mushquash & O'Connor, 2006).

After calculating the variance components in the G-study, we performed D-studies for each EI. A D-study allows a researcher to estimate how a generalizability coefficient would change if the study parameters changed for example by changing the number of students participating in a study or changing the number of items in a factor. We estimated the generalizability coefficients over sample sizes of 25, 50, 75, and 100 students within an institution. By varying the number of students in the D-studies, the results allow us to investigate the dependability of the EIs and describe situations where the use of a group mean is and is not appropriate. We did not calculate generalizability coefficients using different numbers of items as the design of the core NSSE instrument is static.

The study found that the means of the NSSE Engagement Indicators can be reliably generalized to a larger population from small samples of students at postsecondary institutions.

In the D-studies, we calculated two generalizability coefficients using formulas outlined by Kane, Gillmore, and Crooks (1976). We choose to follow Pike's (2006; 2013) approach by calculating both coefficients to obtain more knowledge on the circumstances where using the NSSE Engagement Indicators would be appropriate (see the discussion section for an interpretation of their appropriate uses). The first coefficient generalized over both facets – students and items – and can be interpreted as the expected correlation of the group means derived from two samples of students at the same institution, who answered separate, but comparable items. This generalizability coefficient should be used if a set of items is believed to represent a higher-order construct or a factor. This correlation could also be produced by developing a number of survey items, giving half of the items to half of the students at each institution and correlating the mean to the mean of the other half of items given to the remaining students. The formula used to calculate the coefficient that generalized over students (S) and items (I) was:

$$\epsilon\rho^2(S, I) = \frac{\sigma^2(u)}{\sigma^2(u) + \frac{1}{k}\sigma^2(ui) + \frac{1}{n}\sigma^2(s, ui) + \frac{1}{n \cdot k}\sigma^2(e)}$$

where,

$\sigma^2(u)$ =variance component associated with the institution from the G-study

$\sigma^2(ui)$ =variance component associated with the institution by item interaction from the G-study

$\sigma^2(s, ui)$ =variance component associated with students nested within institutions crossed with items from the G-study

$\sigma^2(e)$ =variance component associated with the error term from the G study

k =number of items in the factor

n =number of students per institution.

The second coefficient generalized only over students by treating the survey items as fixed, rather than random, effects. This coefficient can be interpreted as the expected correlation of the aggregated means of two samples of students who answered the same items. This formula should be used when a conclusion is to be drawn about a set of items, but not a higher-order construct. An analogous method to produce this correlation is to correlate the group means of two samples of students at each institution answering the same items. The formula used to calculate the coefficient that generalized over students was:

$$\epsilon\rho^2(S) = \frac{\sigma^2(u) + \frac{1}{k}\sigma^2(ui)}{\sigma^2(u) + \frac{1}{k}\sigma^2(ui) + \frac{1}{n}\sigma^2(s,ui) + \frac{1}{n \cdot k}\sigma^2(e)}$$

where the variables and variance components are the same as in equation 2.

Limitations

The primary limitation of the study is that it used the institution as the object of measurement. As there is more variability within than between institutions (NSSE, 2008), the results may exhibit a non-trivial difference if the object of measurement utilized was major field or a demographic characteristic. For example, the dependability of QR may be higher when the object of measurement is the group mean of a major field as this measure varies more between majors than it does between institutions (Rocconi, Lambert, McCormick, & Sarraf, 2013). Additionally, the QI item set included a “not applicable” response option that we recoded to missing for this analysis. As “not applicable” is not an ordered response, we were unable to include this response type in our analyses, but excluding students who answered this response could potentially bias the results. Thus, the QI results should be interpreted with caution. We must also note that the generalizability coefficient discussed in this study differs from Brennan and Kane’s (1978) *Index of Dependability*. While the Index of Dependability utilizes GT, the index is designed for mastery tests (of which NSSE is not) and focuses on decisions regarding a cut score such as an admission requirement to score at least 1,000 on the SAT.

Thus, the NSSE EIs can efficiently discriminate institutional environments that promote engagement in effective educational practices. In other words, using a relatively small sample of respondents, the EIs can identify institutions with high and low levels of engagement.

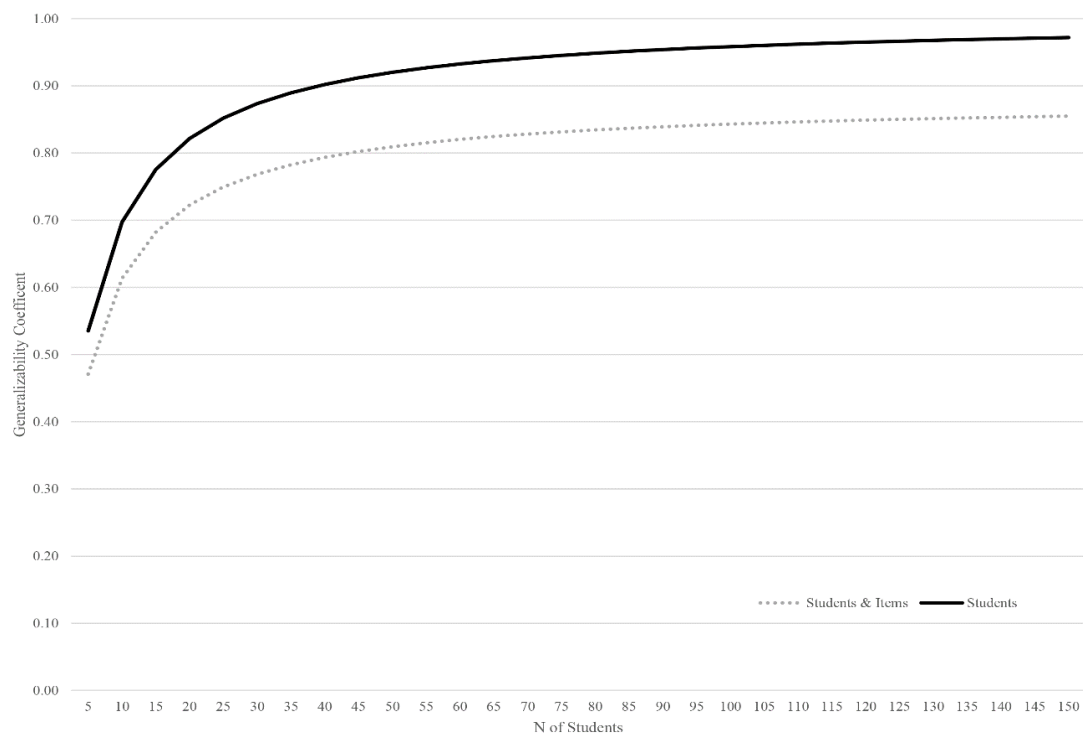
Results

Figure 1 demonstrates the utility of a G- and D-study as it plots both generalizability coefficients by the number of students included in a group mean for the CL engagement indicator for first-year students. The grey dotted line contains the generalizability coefficient of the group mean for various numbers of students when generalizing over students and items. The black line does the same when generalizing over just students. In both lines, there is a steep change in the lines’ slope until an N of roughly 20 students is reached. From about 20 to 60 students, there is a moderate increasing slope, which then flattens for larger numbers of students. Using $E\rho^2 \geq .70$ as a threshold for dependability, at least ten students are required to produce a dependable group mean when generalizing over just students for CL. In contrast, approximately 20 students are required to meet the same threshold when generalizing over students and items. Thus, depending upon the intended purposes of a study, 10 to 20 randomly-selected students would be required to create a dependable group mean that can be generalized to the larger first-year student body for CL. Group means containing fewer students should be viewed as less generalizable and used with caution by researchers or assessment professionals.

Table 1 contains the generalizability coefficients when generalizing over students and items by class and the four sample sizes investigated in the D-studies (the variance components from the G-study are available in Appendix). For first-year students, the CL, DD, SE, and SF EIs met accepted standards for dependability ($E\rho^2 \geq .70$) when a group mean was derived from a sample of 25 to 75 students. The other EIs required substantially larger samples to meet the

Figure 1

D-study generalizability coefficients for first-year students and collaborative learning by number of students

**Table 1**

D-study generalizability coefficients over students and items by class and sample size

	First-Year				Senior			
	25	50	75	100	25	50	75	100
<i>Academic Challenge</i>								
Higher-Order Learning	.48	.56	.59	.61	.46	.54	.57	.58
Reflective & Integrative Learning	.56	.62	.64	.65	.60	.65	.67	.68
Quantitative Reasoning	.39	.48	.52	.55	.46	.55	.59	.61
Learning Strategies	.47	.58	.63	.65	.60	.70	.74	.76
<i>Learning with Peers</i>								
Collaborative Learning	.75	.81	.83	.84	.75	.81	.83	.84
Discussions with Diverse Others	.62	.69	.72	.73	.62	.69	.71	.72
<i>Experiences with Faculty</i>								
Student-Faculty Interaction	.65	.72	.75	.76	.78	.83	.85	.86
Effective Teaching Practices	.47	.55	.58	.59	.50	.57	.60	.61
<i>Campus Environment</i>								
Quality of Interactions	.52	.60	.63	.65	.58	.67	.70	.72
Supportive Environment	.68	.72	.74	.75	.71	.76	.77	.78

same threshold. For seniors, the same EIs had generalizability coefficients greater than .70 using sample sizes of 25 to 50 students. The LS and QI EIs were dependable when a group mean contained at least 75 students. Coefficients for the remaining EIs were below the .70 standard using sample sizes of less than 100 when generalizing over students and items.

The generalizability coefficients when generalizing only over students are located in Table 2. In contrast to the coefficients over both students and items, nearly all of the EIs met standards for dependability using samples as low as 25 students. The exceptions were QR for both classes and LS for first-year students. All the generalizability coefficients were higher than .80 when group means contained 50 seniors, and all were greater than .80 when group means contained 75 first-year students.

Table 2

D-study generalizability coefficients over students but not by items by class and sample size

	First-Year				Senior			
	25	50	75	100	25	50	75	100
<i>Academic Challenge</i>								
Higher-Order Learning	.72	.84	.89	.91	.72	.84	.89	.91
Reflective & Integrative Learning	.81	.90	.93	.95	.85	.92	.94	.96
Quantitative Reasoning	.63	.77	.84	.87	.67	.80	.86	.89
Learning Strategies	.63	.77	.84	.87	.71	.83	.88	.91
<i>Learning with Peers</i>								
Collaborative Learning	.85	.92	.95	.96	.85	.92	.94	.96
Discussions with Diverse Others	.81	.89	.93	.94	.81	.90	.93	.95
<i>Experiences with Faculty</i>								
Student-Faculty Interaction	.80	.89	.92	.94	.88	.93	.96	.97
Effective Teaching Practices	.73	.84	.89	.92	.75	.86	.90	.92
<i>Campus Environment</i>								
Quality of Interactions	.74	.85	.90	.92	.74	.85	.90	.92
Supportive Environment	.87	.93	.95	.96	.89	.94	.96	.97

Discussion

The study found that the means of the NSSE Engagement Indicators can be reliably generalized to a larger population from small samples of students at postsecondary institutions. Therefore, the EIs appear to be dependable measurements of undergraduates' engagement in beneficial activities at an institution during university. Eight of the ten indicators had generalizability coefficients above .70 for both first-year students and seniors, when an institution's mean was derived from just 25 students. All EIs had generalizability coefficients in excess of .70 when the sample size increased to 50 students. Thus, the NSSE EIs can efficiently discriminate institutional environments that promote engagement in effective educational practices. In other words, using a relatively small sample of respondents, the EIs can identify institutions with high and low levels of engagement.

However, the results revealed that only some of the indicators could be dependably generalized to a higher-order construct. The CL, DD SF, and SE EIs appear to be dependable group-level measures when generalizing over students and items and using sample sizes of 25 to 75 students. LS and QI also appear to be dependable for seniors using a sample size of at least 75 students. However, the remaining EIs do not appear to produce dependable group means representing a higher-order construct, except when the sample contains hundreds of students. Therefore, when the object of measurement is an institution, the indicators with lower levels of dependability when generalizing over students and items would be most reliably treated as indexes (groups of items that, when combined, indicate a more general characteristic) rather than higher-order constructs. The lower level of dependability in these indicators, when generalizing over students and items, is generally caused by the small amount of variability accounted for by the institutional effects, which limits the ability to discriminate between institutional means.

Future research should examine the generalizability of NSSE for subgroups (e.g., racial/ethnic groups, major fields, program participants), which will allow users of NSSE data to improve and target their educational offerings.

It is not surprising that some of the indicators have poor dependability as higher-order constructs. NSSE was designed to estimate to undergraduates' engagement in effective educational practices "known to be related to important [university] outcomes" (Kuh et al., 2001, p. 3). Therefore, the Engagement Indicators do not contain items randomly selected from a domain of all possible questions related to a higher-order construct, but rather function as an index or snapshot of the level of engagement in specific beneficial activities known to improve university outcomes. While this study examined the generalizability of the Engagement Indicators over students and items, the purpose of and methods used to construct the survey suggest that this is not the appropriate criterion to assess the dependability of the NSSE Engagement Indicators. Instead, the more appropriate measure is the generalizability coefficient when generalizing over students, but not items.

NSSE briefly examines multiple forms of student engagement to increase its utility for institutions and to ensure a reasonable survey length for respondents. The downside of this approach is that NSSE is unable to ask a detailed set of questions about each type of student engagement. As the accuracy of a student's score is a function of a measurement's reliability (Wainer & Thissen, 1996) and the reliability is related to the number of items in a measurement (Brown, 1910; Spearman, 1910), the relatively small number of items in each Engagement Indicator suggests that an individual's score is associated with a nontrivial amount of error. However, NSSE overcomes this limitation by shifting the object of measurement from an individual student to the group level and aggregating the EIs into group means. Aggregation naturally increases the number of items in a measurement, which results in a higher degree of reliability for institution-level results.

The generalizability or the related concept of reliability in classical test theory does not alone indicate that a measure is valid. Validity is a multifaceted topic that includes construct validity (which this study focuses on), relevance, value implications, and social consequences (Messick, 1989, 1995). Thus, generalizability alone does not indicate that a measure is valid. We encourage readers to also review NSSE's (2018c, 2018d) psychometric portfolio and conceptual framework before concluding that the NSSE Engagement Indicators are accurate measures of student engagement.

Due to the relatively small number of students needed to produce a dependable group mean, the NSSE Engagement Indicators provide the opportunity for assessment professionals to investigate the level of student engagement in a variety of subpopulations.

The vast majority of variation in NSSE data occurs within institutions (NSSE, 2008). In other words, students vary considerably more than institutions. Research and assessment professionals can exploit this variation to examine how a program or academic unit with a high graduation rate impacts students. For example, by comparing the NSSE Engagement Indicators between participants and non-participants in a learning community with a high graduation rate, an institutional researcher may discover that the participants have more academic interactions with their peers and perceive a more supportive campus environment. Administrators may use this finding to justify expanding the program or to implement a portion of the program for all students. Similarly, enrollment in a major may be low because the faculty has poor pedagogical practices that can be improved upon through workshops or another type of intervention. These hypothetical examples illustrate how NSSE data can be used by institutions to identify areas of strength and weakness. After identifying these areas, institutions can intervene to improve areas of weakness and encourage other programs or academic units to adopt the practices of successful programs.

Future research should examine the generalizability of NSSE for subgroups (e.g., racial/ethnic groups, major fields, program participants), which will allow users of NSSE data to improve and target their educational offerings. Alternately, research could examine the generalizability of results of specific types of institutions like publicly controlled colleges and universities or Jesuit colleges. Researchers should also examine the relationship between the NSSE Engagement Indicators and important outcomes like institutional retention rates, completion rates, and student loan default.

In summary, the means of the NSSE Engagement Indicators can be dependably and accurately generalized to a broader population of students when derived from a relatively small sample of undergraduates. The number of students required to produce a dependable group

mean varies by Engagement Indicator; however, a sample of 25 or 50 students is typically sufficient. Due to the relatively small number of students needed to produce a dependable group mean, the NSSE Engagement Indicators provide the opportunity for assessment professionals to investigate the level of student engagement in a variety of subpopulations. Finally, researchers should keep in mind that NSSE is intended to be used as a group-level instrument and was not designed to predict the outcome of an individual student.

References

- Association of American Colleges and Universities. (n.d.). *An introduction to LEAP*. Retrieved from https://www.aacu.org/sites/default/files/files/LEAP/Introduction_to_LEAP.pdf.
- Barr, R. B., & Tagg, J. (1995). From teaching to learning: A new paradigm for undergraduate education *Change*, 27 (November/December), 13-25.
- Brennan, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement*, 14(3), 277-289.
- Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3(3), 296-322.
- Carey, K. (2006). College rankings reformed: The case for a new order in higher education. Washington, DC: Education Sector.
- Carini, R. M., Kuh, G. D., & Klein, S. P. (2006). Student engagement and student learning: Testing the linkages. *Research in Higher Education*, 47(1), 1-32.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.
- Gaston, P. L. (2010). *General education & liberal learning: Principles of effective practice*. Washington, D.C.: Association of American Colleges and Universities.
- Fosnacht, K., Sarraf, S., Howe, E., & Peck, L. K. (2017). How important are high response rates for college surveys?. *The Review of Higher Education*, 40(2), 245-265.
- Hughes, R., & Pace, C. R. (2003). Using NSSE to study student retention and withdrawal. *Assessment Update*, 15(4), 1-2, 15.
- Kane, M. T., & Brennan, R. L. (1977). The generalizability of class means. *Review of Educational Research*, 47(2), 267-292.
- Kane, M. T., Gillmore, G. M., & Crooks, T. J. (1976). Student evaluations of teaching: The generalizability of class means. *Journal of Educational Measurement*, 13(3), 171-183.
- Kirk, R. E. (2013). *Experimental Design: Procedures for the Behavioral Sciences* (4th ed.). Thousand Oaks, CA: Sage.
- Kuh, G. D. (2003). What we're learning about student engagement from NSSE: Benchmarks for effective educational practices. *Change: The Magazine of Higher Learning*, 35(2), 24-32.
- Kuh, G. D. (2008). *High-Impact educational practices: What they are, who has access to them, and why they matter*. Washington, DC: Association of American Colleges and Universities.
- 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., 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., Kinzie, J., Cruce, T., Shoup, R., & Gonyea, R. M. (2006). *Connecting the dots: Multi-faceted analyses of the relationships between student engagement results from the NSSE, and the institutional practices and conditions that foster student success: Final report prepared for Lumina Foundation for Education*. Bloomington, IN: Indiana University, Center for Postsecondary Research.
- Loes, C., Pascarella, E., & Umbach, P. (2012). Effects of diversity experiences on critical thinking skills: Who benefits? *Journal of Higher Education*, 83(1), 1-25.
- Lumina Foundation. (2011). *The Degree Qualifications Profile*. Indianapolis, IN: Author.
- Mayhew, M. J., Seifert, T. A., Pascarella, E. T., Nelson Laird, T. F., & Blaich, C. (2012). Going deep into mechanisms for moral reasoning growth: How deep learning approaches affect moral reasoning development for first-year students. *Research in Higher Education*, 53, 26-46.
- McClenney, K. M., & Marti, C. N. (2006). *Exploring relationships between student engagement and student outcomes in community colleges: Report on validation research*. Retrieved from <http://www.ccsse.org/center/resources/docs/publications/CCSSE Working Paper on Validation Research December 02006.pdf>

- McCormick, A. C., Kinzie, J., & Gonyea, R. M. (2013). Student engagement: Bridging research and practice to improve the quality of undergraduate education. In M. B. Paulsen (Ed.) *Higher education: Handbook of theory and research* (pp. 47-92). Dordrecht, NL: Springer.
- McGuire, M. D. (1995). Validity issues for reputational studies. *New Directions for Institutional Research*, 88, 45-59.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.) *Educational Measurement* (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.
- 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.
- Mushquash, C., & O'Connor, B. A. P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods*, 38(3), 542-547.
- National Survey of Student Engagement. (2008). Promoting engagement for all students: The imperative to look within - 2008 results. Bloomington, IN: Indiana University Center for Postsecondary Research.
- National Survey of Student Engagement. (2013). *NSSE 2013 overview*. Retrieved from http://nsse.indiana.edu/2013_Institutional_Report/pdf/NSSE_2013_Overview.pdf
- National Survey of Student Engagement. (2018a). *Engagement indicators*. Retrieved from http://nsse.indiana.edu/html/engagement_indicators.cfm
- National Survey of Student Engagement. (2018b). *Information about the 2013 update*. Retrieved from <http://nsse.indiana.edu/nsse-update/>
- National Survey of Student Engagement. (2018c). *NSSE's commitment to data quality*. Retrieved from http://nsse.indiana.edu/html/psychometric_portfolio.cfm
- National Survey of Student Engagement. (2018d). *NSSE's conceptual framework (2013)*. Retrieved from http://nsse.indiana.edu/html/conceptual_framework_2.cfm
- Pace, C. R. (1985). *The credibility of student self-reports*. Los Angeles, CA: Center for the Study of Evaluation, University of California.
- Padgett, R. D., Goodman, K. M., Johnson, M. P., Saichaie, K., Umbach, P. D., & Pascarella, E. T. (2010). The impact of college student socialization, social class, and race on need for cognition. *New Directions for Institutional Research*, 145, 99-111.
- Pascarella, E. T., Seifert, T. A., & Blaich, C. (2010). How effective are the NSSE benchmarks in predicting important educational outcomes? *Change*, 42(1), 16-22.
- Pascarella, E. T., & Terenzini, P. T. (2005). *How college affects students: A third decade of research*. San Francisco: Jossey-Bass.
- Pike, G. R. (1994). Applications of generalizability theory in higher education assessment research. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 10, pp. 45-87). New York: Agathon Press.
- Pike, G. R. (2006). The dependability of NSSE Scalelets for college- and department-level assessment. *Research in Higher Education*, 47, 177-195.
- Pike, G. R. (2011). Using college students' self-reported learning outcomes in scholarly research. In S. Herzog & N. A. Bowman (Eds.), *Validity and limitations of college student self-report data*. *New Directions for Institutional Research*, 150, 41-58.
- Pike, G. R. (2013). NSSE benchmarks and institutional outcomes: A note on the importance of considering the intended uses of a measure in validity studies. *Research in Higher Education*, 54(2), 149-170.
- Porter, S. R. (2011). Do college student surveys have any validity? *The Review of Higher Education*, 35(1), 45-76.
- Rocconi, L. M., Lambert, A. D., McCormick, A. C., & Sarraf, S. A. (2013). Making college count: An examination of quantitative reasoning activities in higher education. *Numeracy*, 6(2).
- Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3(3), 271-295.
- Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice*, 15(1), 22-29.

Appendix

Variance components from the G-study by class

Engagement Indicator	$\sigma^2(u)$	$\sigma^2(i)$	$\sigma^2(s,ui)$	$\sigma^2(ui)$	$\sigma^2(e)$	k
<i>First-year</i>						
Higher-Order Learning	0.011	0.000	0.000	0.023	0.642	4
Reflective & Integrative Learning	0.011	0.000	0.000	0.036	0.671	7
Quantitative Reasoning	0.012	0.000	0.000	0.022	0.846	3
Learning Strategies	0.012	0.000	0.000	0.012	0.730	3
Collaborative Learning	0.037	0.000	0.004	0.020	0.715	4
Discussions with Diverse Others	0.026	0.000	0.001	0.030	0.794	4
Student-Faculty Interaction	0.025	0.000	0.000	0.024	0.798	4
Effective Teaching Practices	0.011	0.000	0.024	0.029	0.650	5
Quality of Interactions	0.047	0.000	0.047	0.097	2.683	5
Supportive Environment	0.022	0.000	0.000	0.052	0.839	8
<i>Senior</i>						
Higher-Order Learning	0.011	0.000	0.000	0.024	0.639	4
Reflective & Integrative Learning	0.015	0.000	0.000	0.043	0.659	7
Quantitative Reasoning	0.017	0.000	0.000	0.023	0.897	3
Learning Strategies	0.021	0.000	0.000	0.012	0.779	3
Collaborative Learning	0.043	0.000	0.029	0.023	0.745	4
Discussions with Diverse Others	0.026	0.000	0.000	0.032	0.767	4
Student-Faculty Interaction	0.061	0.000	0.012	0.030	0.910	4
Effective Teaching Practices	0.012	0.000	0.025	0.031	0.654	5
Quality of Interactions	0.059	0.000	0.124	0.083	2.700	5
Supportive Environment	0.030	0.000	0.000	0.059	0.894	8

Note: k = number of items in the Engagement Indicator; All analyses estimated with 50 students