Contextualizing Student Engagement Effect Sizes: An Empirical Analysis

Louis Rocconi and Robert M. Gonyea

Indiana University Bloomington

This paper was presented at the 2015 Association for Institutional Research Conference in Denver, CO on May 28, 2015.

Author Note

Louis Rocconi, Center for Postsecondary Research, School of Education, Indiana University Bloomington; Robert M. Gonyea, Center for Postsecondary Research, School of Education, Indiana University Bloomington. The authors acknowledge the contributions of Pu-Shih Daniel Chen, Counseling and Higher Education, University of North Texas, for analysis on an early version of this paper. Correspondence concerning this article should be addressed to Louis Rocconi, Center for Postsecondary Research, School of Education, Indiana University Bloomington, 1900 East Tenth Street, Suite 419, Bloomington, IN 47406-7512. E-mail: lrocconi@indiana.edu

Abstract

The concept of effect size—a measure of the strength of association between two variables—plays a crucial role in assessment, institutional research, and scholarly inquiry, where it is common with large sample sizes to find small or even trivial relationships or differences that are statistically significant. Using the distributions of effect sizes from the results of 984 institutions that participated in the National Survey of Student Engagement (NSSE) in 2013 and 2014, the authors empirically derived new recommendations for the interpretation of effect sizes which were grounded within the context of the survey. The authors argue for the adoption of new values for interpreting *small*, *medium*, and *large* effect sizes from statistical comparisons of NSSE Engagement Indicators, High-Impact Practices, and student engagement data more generally.

Keywords: effect size, student engagement, institutional assessment

Contextualizing Student Engagement Effect Sizes: An Empirical Analysis

Are your institution's engagement scores are better than those of your comparison group?

Do your students collaborate more in their learning, on average, than their counterparts at similar institutions?

Do your humanities majors engage in a sufficient amount of quantitative reasoning?

Should your institution place more emphasis on High-Impact Practices? Which ones?

To inform answers to questions like these, institutions commonly refer to statistical comparisons from survey results. Yet, simply, knowing that one score is statistically greater than another is not particularly helpful. Statistical significance is often observed at even the most stringent alpha levels (e.g., $\alpha = .001$), especially in research that involves large data sets. It is common with large sample sizes to find small or even trivial relationships or differences that are statistically significant, potentially leading decision makers to redistribute precious resources based on less meaningful effects.

Thus, the concept of *effect size*—a measure of the strength of association between two variables (Grissom & Kim, 2005, 2012)—plays a crucial role in institutional research, assessment, and higher education in general. Jacob Cohen (1969, 1988), one of the most cited experts on the use of effect sizes, is credited with popularizing the term (Kirk, 1996; Huberty, 2002). Cohen (1988) defines an effect size as "the degree to which the phenomenon is present in the population" (p. 9).

Effect sizes are important because they allow researchers to communicate *practical* significance by presenting the magnitude of the effects in standardized metrics and which can be understood regardless of the scale used. Because they are standardized, effect sizes are particularly useful with abstract measurement indices, like those often found in survey research (e.g., NSSE's Engagement Indicators).

Criticisms of hypothesis testing

Null-hypothesis significance testing (NHST) has long been regarded as imperfect for examining data (Cohen, 1994; Kirk, 1996; Kline, 2013; Hill & Thompson, 2004). Tests of statistical significance provide insight into whether the observed differences might have occurred by chance alone. However,

scholars (Cohen, 1994; Kirk, 1996; Kline, 2013; Hill & Thompson, 2004) have noted numerous problems with NSHT. We summarize three main criticisms. The first criticism concerns a misunderstanding of *p*-values. In NSHT, the *p*-value tells us the probability of obtaining these data or more extreme data (D) given that the null hypothesis (Ho) is true, that is p(D|Ho). However, researchers sometimes misinterpret the *p*-value from statistical tests (Cohen, 1994; Kirk, 1996; Lipsey et al., 2012; Kline, 2013) to mean the probability the null hypothesis is true given that we have observed these data, that is *p*(Ho|D). Unfortunately for researchers *p*(D|Ho) ≠ *p*(Ho|D); nor does obtaining data with a small *p*(D|Ho) imply that *p*(Ho|D) is also small (Cohen, 1994; Kirk, 1996). A second criticism is that NSHT evaluates sample size. Given a large enough sample size, any statistic can be found to be statistically significant. As Thompson (1998) jokingly stated, "If we fail to reject, it is only because we've been too lazy to drag in enough participants" (p. 799). A third criticism is that statistical significance does not equal practical significance. Statistical significance evaluates the probability of sample results; however, these tests of statistical significance do not provide insight into whether the magnitude of these effects are substantively important – an issue of particular interest to policy makers. Statistical significance merely means statistical rareness, but unlikely events can be completely meaningless or trivial, or conversely, likely events may nevertheless be quite noteworthy. *p*-values are not good judges of practical importance since they are confounded by the joint influence of sample results and sample size.

Types of effect sizes

While in this paper we will discuss two specific effect sizes, Cohen's *d* and Cohen's *h*, there are in fact numerous effect size statistics. Kirk (1996) and Rosnow and Rosenthal (2003) classify effect sizes into three broad categories of measures. While these scholars give different names to the categories, they generally are described as (a) measures of difference, (b) measures of strength of association, and (c) other measures. Measures of difference are sometimes referred to as the *d*-type family of effect sizes, after Cohen's popular *d* statistic. These effect sizes measure the magnitude of the distance between groups, and include raw differences (e.g., $M_1 - M_2$), standardized differences (e.g., Cohen's *d*, Glass's *g*), and transformed differences (e.g., Cohen's *h*, Cohen's *q*, probit *d*). Measures of strength of association are

also known as the *r*-type family of effect sizes after the popular Pearson's product-moment correlation coefficient – the *r* statistic. This family of measures is concerned with measures of correlation and variance accounted for and includes such statistics as a correlation, r, and r^2 and eta-squared (μ^2) and omega-squared (ω^2) from ANOVA. The third category includes other measures of effect such as the odds ratio or relative risk.

Cohen's *d* effect size is used to describe the standardized mean difference between two groups of independent observations. It is calculated by dividing the mean difference by the pooled standard deviation. While it was Hedges (1982) who proposed using the pooled sample standard deviation to standardize the mean difference, we will continue to refer to this effect size by its more common name of Cohen's *d* (Fritz, Morris, & Richler, 2011). The formula to compute Cohen's *d* is as follows: $d = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{(n_1-1){s_1}^2+(n_2-1){s_2}^2}{n_1+n_2-2}}}$.

Cohen's *h* effect size is the difference between two independent group proportions after each proportion has been transformed using an arcsine transformation. Specifically, it is calculated as follows: $h = (2\sin^{-1}\sqrt{\mathrm{P}_1}) - (2\sin^{-1}\sqrt{\mathrm{P}_2})$. The reason for employing the arcsine transformation is to make the proportions comparable in the sense of having variances independent of the parameter (Cohen, 1988; Rosnow & Rosenthal, 2003; Hojat & Xy, 2004). This type of transformation is known as a variance stabilizing transformation. For instance, the variance of a proportion is equal to the proportion multiplied by one minus the proportion divided by the sample size [$(p) = \frac{(p)(1-p)}{n}$, where *p* represents the proportion]. Thus, the variance of a proportion is dependent upon the value of the proportion. The fact that the variance of the proportion depends on its particular value prevents the simple difference between proportions to be used in power calculations because constant differences between two proportions cannot always be considered equal on the scale of proportions (Cohen, 1988). It is easier to detect differences between proportions that fall on the ends of the proportion scale than it is to detect differences between proportions that fall in the middle of the proportion scale (in other words, it gets easier to detect

differences the further a proportions falls from .5). Thus, a transformation must be made to the proportions such that differences between the transformed parameters are equally detectable. Values for Cohen's *h* range from $-\pi$ to π, or around -3.14 to 3.14 this is because values of the arsine function range between $-\pi/2$ and $\pi/2$.

Interpreting effect sizes

Cohen (1988, 1992) described *small* effects as those that are hardly visible, *medium* effects as observable and noticeable to the eye of the beholder, and *large* effects as plainly evident or obvious. He then reluctantly suggested that *d* values (and *h* values) of about .2, .5, and .8, and *r* values of about .1, .3, and .5, would be small, medium, and large effects respectively. Yet, Cohen cautioned that "there is a certain risk inherent in offering conventional operational definitions for those terms for use in power analysis in as diverse a field of inquiry as behavioral science" (p. 25) and urged researchers to interpret effect sizes within the *context of the data*. Nevertheless, Cohen's recommendation has been incorporated into many social science studies.

Cohen (1988) also cautioned that when a phenomenon cannot be brought into the laboratory to be studied, which is the case in the vast majority of higher education research, extraneous or uncontrollable factors lead to smaller or more difficult-to-detect effect sizes In fact, in the realm of social science and educational research, Cohen was right. Many scholars in those fields have noted that study effects were often small by Cohen's standards (Rosnow & Rosenthal, 2003; Valentine & Cooper, 2003; Lipsey et al., 2012), leading to difficulties in interpretation (Ferguson, 2009). For example, Hill, Bloom, Black, and Lipsey (2008) summarized estimates of achievement effect sizes from random assignment studies of K-12 educational interventions and noted that the mean effect sizes typically ranged from .20 to .3. Similarly, investigating K-12 students' academic performance on standardized reading and mathematics achievement tests, Lipsey et al. (2012) found effect sizes to rarely be as large as .30 and noticed a relatively consistent pattern of smaller effect sizes for high school students. In another example, a meta-analysis of 62 studies investigating the impact of service-learning on P-16 student outcomes, Celio, Durlak, & Dymnicki (2011) found average effect sizes ranging from .27 to .43. In the business field, Ellis

(2010) summarized that around two-thirds of effect sizes reported in international business were small by Cohen's standards. Finally, in medicine, McCartney & Rosenthal (2000) noted that in research involving hard to change outcomes, such as the incidence of heart attacks, the largest effect size found was below .20, what Cohen considered small. However, those small effects corresponded to reducing the incidence heart attacks by about half, an enormous practical significance.

While there is usually no agreement on the size an effect must be to constitute practical significance – it is likely to depend on the context of each study. Hill and colleagues note that effect size estimates were much lower when the outcome measure was a standardized test which covered a broad subject matter such as the SAT9 composite reading test (.07), than when the outcome focused on a specialized topic such as reading comprehension developed by the researcher for that intervention (.44). Lipsey et al. (2012) noticed a distinctive pattern in the average effect sizes found for different kinds of achievement test measures (e.g., whether they were specialized or broadly focused).

Hill *et al* (2008) summed it up: "Empirical benchmarks from a research synthesis do not indicate what effects are *desirable* from a policy standpoint. Instead, they provide a snapshot of effects found in previous studies, that is, what might be *attainable*" (p. 176).

Further complicating the interpretations of effect sizes, Cohen's own recommendations are not even consistent across different effect size types. For example, Cohen suggested that both $d = .5$ and $r = .3$ indicate a **medium** effect size. Yet, using the formula $r = d/\sqrt{(d\text{^}2+4)}$, we know that $d = .5$ is the equivalent of $r = .24$, which would be considered a **small** effect by *r* standards. Similarly, a **large** *d* effect of .8 corresponds to $r = .37$, just over the **medium** threshold for an *r* effect. Thompson (2001) noted that if researchers interpret effect sizes using fixed benchmarks with the same rigidity as the $p = .05$ has been used in statistical testing, "we would merely be being stupid in another metric" (p. 83). Researchers and administrators assessing student engagement need the ability to interpret effect sizes of their results—whether against a comparison group of institutions or between subgroups of their own students.

Nevertheless, the American Psychological Association's (APA) publication manual is clear about the importance of reporting effect sizes: "For the reader to appreciate the magnitude or importance of a

study's findings, it is almost always necessary to include some measure of effect size" (APA, 2010, p. 34). Additionally, the APA Task Force emphasized that reporting and interpreting effect sizes with consideration to effects from previously studies is essential to good research (Wilkinson & APA Task Force on Statistical Inference, 1999). Similarly, the American Educational Research Association (AERA, 2006) recommended in its standards for reporting social science research that statistical results be accompanied by an effect size and a "qualitative interpretation" of the effect. Ellis (2010) echoes the stances of the APA and AERA noting that effect sizes are meaningless unless they can be contextualized against some frame of reference.

Purpose and Research Questions

The purpose of this study is to examine the distribution of statistical comparisons and their effects between institutions and their comparison groups using measures from the National Survey of Student Engagement (NSSE), and to make recommendations for the interpretation of effect sizes from engagement results. Therefore, the following research questions guided our study.

1. How do the effect sizes from NSSE institutional comparisons distribute within Cohen's small, medium, and large ranges?
2. Is it possible to derive more useful effect size cut points that fit the context of institutional engagement results?

Methods

Data Source

Data for this study came from the 2013 and 2014 administrations of the National Survey of Student Engagement (NSSE). NSSE is an annual survey administered to first-year and senior students at bachelor's degree-granting colleges and universities across the United States. NSSE is used to assess the extent to which students are exposed to and participate in a variety of effective educational practices (McCormick, Kinzie, & Gonyea, 2013). The survey asks students about various aspects of their undergraduate experience, such as the time and effort they invest in

their studies, their discussions and interactions with students who are different from themselves, their interactions with faculty members and students, and other educationally purposeful activities. The analytic sample consisted of 984 institutions that participated in the 2013 or 2014 administration of NSSE. For institutions that participated both years, we only included their most recent year of participation. Participating institutions represented a broad cross-section of the national profile of U.S. baccalaureate institutions (Table 1). In this study, we used the distribution of effect sizes from NSSE to empirically derive new recommendations for their interpretation.

Measures

Effect sizes for the study were based on comparisons of two primary sets of variables generated from NSSE questionnaire: Engagement Indicators (EIs) and High-Impact Practices (HIPs). NSSE's ten EIs represent the multi-dimensional nature of student engagement, organized within four engagement themes. They include four measures of academic challenge: *Higher-Order Learning, Reflective & Integrative Learning, Learning Strategies,* and *Quantitative Reasoning;* two measures about learning with peers: *Collaborative Learning* and *Discussions with Diverse Others;* two measures describing experiences with faculty: *Student-Faculty Interaction* and *Effective Teaching Practices;* and two measures of the campus environment: *Quality of Interactions* and *Supportive Environment.* Each EI is a reliable scale that measures a distinct aspect of student engagement by summarizing students' responses to a set of related survey questions.

HIPs have positive associations with student learning and retention because they often demand considerable time and effort, facilitate learning outside of the classroom, require meaningful interactions with faculty and students, encourage collaboration with diverse others, and provide frequent and substantive feedback (Kuh, 2008). NSSE asks students if they have participated in six HIPs: learning communities, service-learning, research with a faculty member, internship or field experiences, study

abroad, and culminating senior experiences. (The first three are asked of first-year students, and all six are asked of seniors.)

Analysis

To answer the first research question, we generated a data set by calculating effect sizes for each EI and HIP, separately for first-year and senior students, for each of the 984 institutions compared with respondents from all other institutions in data. Results were weighted by institution-reported sex, enrollment status, and institution size.

To answer the second research question, we considered Cohen's (1988) rationale for observing small, medium, and large effects, and ways in which institutional differences would be observable in the data. First, we assigned percentile rankings to institutions' precision-weighted Engagement Indicator (EI) scores and High-Impact Practice (HIP) scores, and used these percentile rankings to model comparisons that would resemble effect sizes of increasing magnitude (illustrated in Figure 1). We conceptualized that a *small* effect would resemble the difference between the scores of students attending institutions in the third quartile (i.e., between the 50^{th} and 75^{th} percentiles) and those attending institutions in the second quartile (i.e., between the 25^{th} and 50^{th} percentile). These two sets of institutions are labeled groups A and B in Figure 1a. Because groups A and B are fairly close within the distribution, the difference between the students attending those institutions is expected to be small. In a similar way, a *medium* effect would resemble the difference between students attending institutions in the upper and lower halves of the distribution, and a *large* effect would resemble the difference between students attending institutions in the top and bottom quartiles. Finally, we calculated confidence intervals for the effect sizes by bootstrapping 1,000 samples for each comparison group which was used in each effect size calculation.

Results

Research Question 1: How do the effect sizes from NSSE institutional comparisons distribute within Cohen's small, medium, and large ranges?

Table 2 shows the percentage of institutions that had effect sizes within each of Cohen's ranges on the EIs and HIPs for first-year and senior students. The vast majority of effect sizes were either *trivial* (ES < |.2| in magnitude) or *small* (|.2| ≤ ES < |.5|). For most EIs, over 60% of the effect sizes were trivial and over 20% were small. Very few institutions found medium or large EI effect sizes using Cohen's criteria. An exception was Student-Faculty Interaction for seniors, where fewer (41%) effect sizes were classified as trivial, and more were classified as medium (16%).

HIP comparisons showed somewhat different patterns. While the largest number of HIP effect sizes were trivial in magnitude, they ranged between 36% and 84%. Compared to the EIs, more HIP effect sizes were in the medium and large range, particularly among seniors. For example, 17% of first-year effect sizes, and 18% of senior effect sizes for service-learning were at least medium in magnitude. Similar totals were tallied for senior internships and study abroad, and fully 27% of effect sizes for the senior capstone experience were at least medium in magnitude.

Research Question 2: Is it possible to derive more useful effect size cut points that fit the context of institutional engagement results?

Given the fact that a large majority of effect sizes were small or trivial according to Cohen's cut points, we analyzed effect sizes according to the proposed scheme based on the distribution of institutional scores. Table 3 shows the effect sizes and confidence intervals for the small, medium, and large model comparisons for first-year students and seniors on all ten of the NSSE EIs, and Table 4 shows the effect sizes and confidence intervals for these contrived model comparisons on the six HIPs. While

the effect sizes in Table 3 varied somewhat between EIs and between student class levels, the ranges within the small, medium, and large categories were fairly consistent and, with the exception of Student-Faculty Interaction for seniors, did not overlap. That is, the maximum small effect size was almost always lower than the minimum medium effect size, and the maximum medium effect size was lower than the minimum large effect size. For both first-year students and seniors, the average small effect size was about .1 and the average medium effect size was about .3. The average large effect size for first-year students was about .44 and for seniors was about .48. Compared to Cohen's recommendations, these effect sizes were lower and did not range as widely.

Effect sizes varied more across HIPs (Table 4) and across class year than did the EIs. While the effect sizes for learning community and research with faculty were generally similar to those of the EIs, the effect sizes for service-learning, internship, study abroad, and senior capstone were considerably larger and in fact closely approximated Cohen's standards of .2, .5, and .8. Of the three HIPs measured for first-year students, service-learning had the widest range, with small, medium, and large estimates of .18, .43, and .73. On the other hand, research with faculty estimates for first-year students were a bit smaller and in a fairly narrow range, with estimates of .06, .17, and .26 respectively. Still, the average effect size for the three HIPs were .11, .31, and .50 for first-year students – which are consistent with the EI averages. Senior estimates for HIP effect sizes were generally larger, and ranged more. Small, medium, and large average effect sizes for seniors were .18, .46, and .70 respectively. These values look more like Cohen's recommended values of .2, .5, and .8. In sum, comparing these with Cohen's recommendations, it was clear that even though some of the HIPs were better aligned with Cohen, some adjustments were needed.

Discussion

Comparing Table 2 with Tables 2 and 3 suggests that new effect size criteria for the interpretation of EI comparisons is necessary, while Cohen's recommended values adequately fit the distributions of effect sizes from comparisons of the High-Impact Practice measures.

The consistency of effect size values among the Engagement Indicators points toward a new set of criteria for their interpretation: small effects start at about .1, medium effects start at about .3, and large effects start at about .5 (Table 5). These new reference values were selected after an examination of the minimum values in Table 3, which when rounded to the nearest tenth approximated evenly-spaced intervals between .1 and .5. Like Cohen's, these new values should not be interpreted as precise cut points, but rather are to be viewed as a coarse set of thresholds or minimum values by which one might consider the magnitude of an effect size. The simplicity of the proposed values for Engagement Indicators may have intuitive and functional appeal for users of NSSE data. More institutions with real differences will find effect sizes of .3 or .5, and should interpret them as medium or large effects. Furthermore, institutions with effect sizes of .1, although still relatively small, no longer need to disregard them as trivial.

Table 6 reports the distribution of effect sizes based on these proposed reference values. As expected from our previous analysis of effect size distribution, the majority of effect sizes were trivial, small, or medium. Yet, this is a finer distribution within categories from what we saw in Table 2 based on Cohen's definitions. For the EIs, Table 6 shows that approximately 35-40% of all effect sizes are in the trivial range with another 40-45% considered small, and the medium captures about 10-15% of all effect sizes while large effect sizes are relatively rare.

Conclusion

The purpose of this study was to examine the distribution of effect sizes from the National Survey of Student Engagement, and to make recommendations for the interpretation of effect sizes from engagement results. Our analyses informed the development of a new set of reference values for interpreting the Engagement Indicator effect sizes. As a practical matter, at least four approaches can be taken with regard to effect sizes in the context of NSSE and student engagement results. First, it's not unreasonable to continue using Cohen's purposefully vague definition, particularly for high-impact practices. The new reference values offered in Table 5

only deviate slightly from Cohen. Second, for those willing to consider the new reference values proposed in Table 5, the thresholds of .1, .3, and .5 could have appeal for their simplicity and functionality. They are grounded in actual NSSE data and may allow for richer interpretations of NSSE results. Third, it's also possible to ignore the new reference values and to examine the results in Tables 2 and 3 for a more nuanced interpretation of a particular effect size. Tables 2 and 3 reveal a different pattern of effect sizes for each engagement indicator and high-impact practice. What's more, effect sizes for the Student Faculty Interaction and high-impact practices, particularly service-learning, internship, study abroad, and capstone, tend to be larger in magnitude than for other engagement indicators. Finally, we also recommend an examination of individual item frequencies in combination with effect size interpretation. Individual items provide a richer explanation for the magnitude of the effect sizes, and can help administrators and policy makers interpret results in ways that are context-specific and actionable. Be aware that many combinations of individual item results can produce a particular effect size. For example, consider two institutions with the same effect size on a particular engagement indicator. The first may have large percentage differences on just a few of the items and no differences on the others, while the second could have small percentage differences on all the items. Whatever the approach, effect sizes can be a useful statistic to help institutions interpret the strength or magnitude of their engagement indicator and high-impact practice scores in relation to their selected comparison groups.

References

Cohen, J. (1969). Statistical power analysis for the behavioral sciences. New York: Academic Press.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112(1), 155-159.

Cohen, J. (1994). The earth is round ($p<.05$). American Psychologist, 49(12), 997-1003.

Ellis, P. D. (2010). Effect sizes and the interpretation of research results in international business. Journal of International Business Studies, 47, 1581-1588.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. Professional Psychology: Research and Practice, 40(5), 532-538.

Grissom, R. J., & Kim, J. J. (2005). Effect sizes for research: A broad practical approach. Mahwah, NJ: Lawrence Erlbaum Associates.

Grissom, R. J., & Kim, J. J. (2012). Effect sizes for research: Univariate and Multivariate Applications, Second Edition. New York: Routledge.

Hill, C. J., Bloom, H. S., Black, A. B., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. Child Development Perspectives, 2(3), 172-177.

Hill, C. R., & Thompson, B. (2004). Computing and interpreting effect sizes. In J. C. Smart (Ed.). Higher Education: Handbook of Theory and Research, (Vol. 19, pp. 175-195). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Huberty, C. J. (2002). A history of effect size indices. Educational and Psychological Measurement, 62, 227-24.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56(5), 746-759.

Kline, R. B. (2013). Beyond significance testing: Statistics reform in the behavioral sciences (2nd edition). Washington, DC: American Psychological Association.

Lipsey, M., Puzio, K., Yun, C., Hebert, M., Steinka-Fry, K., Cole, M., Roberts, M., Anthony, K., & Busick, M.D. (2012). Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms. (NCSER 2013-3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

McCartney, K., and Rosenthal, R. (2000). Effect size, practical importance, and social policy for children. Child Development, 71(1): 173-18.

McCormick, A. C., J. Kinzie, and R. M. Gonyea. (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, (Vol. 28, pp. 47–92). Dordrecht, The Netherlands: Springer.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44, 1276-1284.

Rosnow, R. L., & Rosenthal, R. (2003). Effect sizes for experimenting psychologists. Canadian Journal of Experimental Psychology, 57(3), 221-237.

Thompson, B. (1998). In praise of brilliance: Where the praise really belongs. American Psychologist, 53, 799-80.

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. Journal of Experimental Education, 70, 80-93.

Valentine, J., & Cooper, H. (2003). Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes. Washington, DC: What Works Clearinghouse.

Tables and Figures

Table 1.
Characteristics of Participating Institutions (N=984)

		%
2010 Basic Carnegie Classification	Research Universities (very high research activity)	5
	Research Universities (high research activity)	7
	Doctoral/Research Universities	6
	Master's Colleges and Universities (larger programs)	27
	Master's Colleges and Universities (medium programs)	11
	Master's Colleges and Universities (smaller programs)	6
	Baccalaureate Colleges--Arts & Sciences	16
	Baccalaureate Colleges--Diverse Fields	17
	Other types	6
Control	Public	40
	Private	60
Barron's Selectivity	Noncompetitive	4
	Less Competitive	10
	Competitive	46
	Very Competitive	19
	Highly Competitive	8
	Most Competitive	3
	Not available/Special	10

Table 2

Frequency of NSSE Effect Sizes by Cohen's Suggested Ranges[a]

	Effect Size Range							
	Trivial		Small		Medium		Large	
	ES < |.2|		|.2| ≤ ES < |.5|		|.5| ≤ ES < |.8|		ES ≥ |.8|	
Engagement Indicator	First-year	Senior	First-year	Senior	First-year	Senior	First-year	Senior
Higher-Order Learning	72%	75%	26%	23%	1%	1%	<1%	<1%
Reflective & Integrative Learning	71%	68%	26%	28%	2%	3%	<1%	1%
Learning Strategies	75%	66%	22%	33%	2%	1%	<1%	<1%
Quantitative Reasoning	76%	79%	20%	18%	2%	2%	1%	<1%
Collaborative Learning	64%	58%	30%	35%	4%	5%	2%	2%
Discussions with Diverse Others	61%	63%	34%	33%	4%	3%	<1%	1%
Student-Faculty Interaction	60%	41%	33%	39%	6%	16%	1%	4%
Effective Teaching Practices	68%	71%	30%	27%	1%	2%	<1%	<1%
Quality of Interactions	59%	59%	37%	37%	2%	4%	<1%	0%
Supportive Environment	61%	55%	34%	38%	4%	6%	<1%	<1%
High-Impact Practice								
Learning Community	57%	69%	38%	26%	3%	3%	1%	1%
Service-Learning	47%	46%	36%	36%	11%	13%	6%	5%
Research with Faculty	84%	55%	15%	32%	1%	11%	0%	2%
Internship[b]	--	43%	--	38%	--	15%	--	4%
Study Abroad[b]	--	40%	--	43%	--	10%	--	7%
Senior Capstone[b]	--	36%	--	36%	--	17%	--	10%

a. Cohen's suggestions of small (d & h = .2), medium (d & h = .5), and large (d & h = .8)

b. Effect sizes are for Internship, Study Abroad, and Senior Capstone are not calculated for first-year students since these opportunities are typically not available to first-year students.

Table 3

Effect Sizes from NSSE EI Percentile Group Comparisons

(95% confidence intervals given in parentheses)

	First-year			Senior		
	Small	Medium	Large	Small	Medium	Large
Higher-Order Learning	.087 (.074, .098)	.223 (.214, .232)	.372 (.359, .385)	.096 (.085, .106)	.246 (.239, .253)	.356 (.346, .365)
Reflective & Integrative Learning	.109 (.098, .121)	.260 (.251, .268)	.394 (.381, .407)	.103 (.094, .113)	.266 (.260, .272)	.414 (.404, .424)
Learning Strategies	.088 (.076, .099)	.227 (.218, .235)	.355 (.342, .368)	.078 (.068, .087)	.203 (.196, .209)	.312 (.302, .322)
Quantitative Reasoning	.092 (.079, .105)	.237 (.229, .246)	.354 (.341, .366)	.113 (.104, .123)	.304 (.298, .312)	.466 (.456, .476)
Collaborative Learning	.129 (.117, .141)	.363 (.354, .371)	.549 (.537, .561)	.125 (.116, .134)	.381 (.375, .388)	.594 (.584, .604)
Discussions with Diverse Others	.133 (.121, .146)	.330 (.321, .339)	.501 (.488, .515)	.120 (.110, .130)	.321 (.314, .329)	.510 (.500, .520)
Student-Faculty Interaction	.121 (.110, .133)	.335 (.326, .344)	.545 (.530, .560)	.194 (.183, .205)	.491 (.483, .498)	.744 (.732, .756)
Effective Teaching Practices	.100 (.087, .112)	.276 (.266, .285)	.414 (.401, .428)	.086 (.076, .096)	.245 (.238, .252)	.373 (.363, .383)
Quality of Interactions	.139 (.127, .152)	.317 (.308, .326)	.461 (.449, .472)	.135 (.124, .146)	.360 (.353, .367)	.515 (.505, .525)
Supportive Environment	.116 (.104, .130)	.310 (.301, .319)	.488 (.475, .501)	.136 (.125, .146)	.344 (.336, .351)	.529 (.519, .540)
Minimum *d*	.087	.223	.354	.078	.203	.312
Maximum *d*	.139	.363	.549	.194	.491	.744
Average *d*	.111	.288	.443	.118	.316	.481

Table 4

Effect Sizes from NSSE High-Impact Practices Percentile Group Comparisons (95% confidence intervals given in parentheses)

	First-year			Senior		
	Small	Medium	Large	Small	Medium	Large
Learning Community	.105 (.093, .118)	.345 (.337, .354)	.513 (.501, .525)	.096 (.086, .107)	.286 (.279, .293)	.434 (.424, .445)
Service-Learning	.179 (.166, .192)	.427 (.419, .437)	.728 (.714, .741)	.171 (.161, .182)	.434 (.427, .441)	.690 (.677, .702)
Research with Faculty	.058 (.045, .070)	.166 (.158, .175)	.255 (.242, .267)	.156 (.146, .165)	.407 (.400, .415)	.606 (.595, .616)
Internship[a]	--	--	--	.199 (.190, .208)	.501 (.494, .508)	.757 (.746, .768)
Study Abroad[a]	--	--	--	.199 (.189, .208)	.499 (.492, .506)	.784 (.775, .793)
Senior Capstone[a]	--	--	--	.246 (.236, .257)	.604 (.596, .612)	.920 (.909, .931)
Minimum *h*	.058	.166	.255	.096	.286	.434
Maximum *h*	.179	.427	.728	.246	.604	.920
Average *h*	.114	.313	.498	.178	.455	.698

a. Effect sizes for Internship, Study Abroad, and Senior Capstone were not calculated for first-year students since these opportunities are typically not available to first-year students.

Table 5

Recommendations for NSSE Effect Size Interpretations

	Effect Size Values	
	Engagement Indicator Comparisons (*d*)	High-Impact Practice Comparisons (*h*)*
Small	≥ .1	≥ .2
Medium	≥ .3	≥ .5
Large	≥ .5	≥ .8

* Particularly for Service-Learning, Internship, Study Abroad, and Capstone

Table 6

Frequency of NSSE Effect Sizes by Suggested Ranges[a]

	Effect Size Range							
	Trivial		Small		Medium		Large	
	ES < |.1|		|.1| ≤ ES < |.3|		|.3| ≤ ES < |.5|		ES ≥ |.5|	
Engagement Indicator	First-year	Senior	First-year	Senior	First-year	Senior	First-year	Senior
Higher-Order Learning	45%	46%	44%	45%	9%	8%	1%	1%
Reflective & Integrative Learning	40%	40%	47%	44%	11%	12%	2%	4%
Learning Strategies	44%	38%	46%	46%	8%	15%	2%	1%
Quantitative Reasoning	47%	49%	42%	41%	8%	7%	3%	3%
Collaborative Learning	34%	30%	46%	48%	14%	14%	5%	7%
Discussions with Diverse Others	33%	35%	47%	47%	15%	14%	4%	4%
Student-Faculty Interaction	33%	23%	43%	34%	17%	23%	6%	20%
Effective Teaching Practices	38%	41%	48%	46%	12%	11%	1%	2%
Quality of Interactions	34%	30%	46%	48%	16%	18%	3%	4%
Supportive Environment	36%	30%	45%	46%	15%	18%	4%	6%

a. Modified effect size ranges of small ($d \geq .1$), medium ($d \geq .3$), and large ($d \geq .5$)

b. Effect sizes are for Internship, Study Abroad, and Senior Capstone are not calculated for first-year students since these opportunities are typically not available to first-year students.

Figure 1

Illustration of Four Model Comparison Groups for Determining Empirically-Based Effect Size Thresholds Based on the Distribution of Student Engagement Measures