
## Abstract

The concept of effect size plays a crucial role in assessment, institutional research, and scholarly inquiry, where it is common with large sample sizes to find small relationships that are statistically significant. This study examines the distribution of effect sizes from institutions that participated in the National Survey of Student Engagement (NSSE) and empirically derives recommendations for their interpretation. The aim is to provide guidelines for researchers, policymakers, and assessment professionals to judge the importance of an effect from student engagement results. The authors argue for the adoption of the recommendations for interpreting effect sizes from statistical comparisons of NSSE data.



## AUTHORS

Louis M. Rocconi, Ph.D.
*University of Tennessee, Knoxville*

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


# Contextualizing Effect Sizes in the National Survey of Student Engagement: An Empirical Analysis

The concept of effect size plays a crucial role in higher education assessment. Assessment professionals tasked with gauging the success of campus policies and interventions often use effect sizes of their most important outcome measures (e.g., Springer, 2006). Many of these efforts rely on statistical comparisons where stakeholders not only want to know whether an intervention or policy has an effect, but also *how large the effect is*. Simply knowing that one score is statistically different from another is not particularly helpful. Especially in research that involves large data sets, it is common to find very small relationships or differences that are statistically significant at even the most stringent alpha levels (e.g., $\alpha = .001$). This could lead decision-makers to redistribute precious resources based on matters that are immaterial. On the other hand, decisions may be better informed if based on the relative magnitude of the effect. Thus, estimates of effect size provide researchers and practitioners essential information on the practical or theoretical importance of research findings. However, to better interpret the substantive value of an effect, effect sizes need to be grounded within a meaningful context.

The aim of this article is to examine the distribution of effect sizes from institutional comparisons reported by the National Survey of Student Engagement (NSSE) and make recommendations for their interpretation. We begin with an introduction to NSSE and its use in higher education assessment. Next, we provide a definition of effect size and a review of the limitations of hypothesis testing. We then discuss different types of effect sizes and the challenges involved in interpreting them in different contexts. Then, after considering Cohen's (1988) rationale for interpreting the size of an effect, we use the distribution of NSSE effect sizes from nearly a thousand participating institutions as a normative context to interpret the "natural" or relative variation in magnitudes of institution-to-peer-group comparisons. Ultimately, our aim is to provide helpful guidelines for assessment professionals, policymakers, and researchers to judge the importance of their student engagement results.


## CORRESPONDENCE

***Email***
lrocconi@utk.edu

## Background: The National Survey of Student Engagement

NSSE is an annual survey administered to first-year and senior students at bachelor's degree-granting colleges and universities across the United States and Canada. NSSE is used to assess the extent to which undergraduate students are exposed to and participate in a variety of effective educational practices (McCormick, Kinzie, & Gonyea, 2013). Decades of research on undergraduate students (see Astin, 1993; McCormick et al., 2013; Pace, 1979; Pascarella & Terenzini, 1991, 2005) show that students benefit from college when their efforts are directed at learning-centered activities both inside and outside of the classroom. In an effort to leverage these ideas to inform the assessment and improvement of undergraduate education, the National Survey of Student Engagement was launched in 2000. Standardized sampling and administration procedures ensure the comparability of results among participating institutions.

**Thus, estimates of effect size provide researchers and practitioners essential information on the practical or theoretical importance of research findings. However, to better interpret the substantive value of an effect, effect sizes need to be grounded within a meaningful context.**

Since its launch in 2000, NSSE has been used in institutional assessment as a valid source of evidence, whether by itself or linked with other school records (see McCormick et al., 2013 for a review). Colleges and universities participate in NSSE for a variety of reasons but mainly to assess the quality of their curricular and co-curricular undergraduate learning programs. As such, NSSE provides a suite of student engagement measures—including 10 Engagement Indicators, six High-Impact Practices, and items about the amount time spent preparing for classes, the quantity of reading and writing, perceived course challenge, and more. NSSE content can be mapped to department, institution, or accreditation goals and can be used to evaluate key performance indicators or to track progress on a strategic plan. NSSE also provides comparative data on these measures from other participating campuses (in aggregate). Such comparisons are valuable to know where to direct institutional improvement efforts. Effect sizes from these comparisons are used to identify dimensions of student learning where the institution is doing well, and areas where improvement is warranted (for a discussion of using effect sizes in NSSE reporting see Springer, 2006). The NSSE website (nsse.indiana.edu) and their *Lessons from the Field* series (NSSE, 2015, 2017) catalog hundreds of examples of how colleges and universities employ engagement data in this way. In many of these examples, effect sizes provide a way not only to identify meaningful differences between the institution and comparison group but also to track the magnitude of changes across multiple years of NSSE administrations on the same campus.

## Definition of Effect Size

While Jacob Cohen (1988, 1992) is credited with popularizing the use of effect sizes, the idea of supplementing significance tests with an effect size statistic can be traced back to the early 1900s and the works of Karl Pearson and Ronald Fisher (Fisher, 1925; Pearson, 1900). Cohen (1988) defines an effect size as "the degree to which the phenomenon is present in the population" (p. 9). Effect sizes have also been described as the degree to which results differ from the null hypothesis (Grissom & Kim, 2005, 2012), the degree to which study results should be considered important regardless of sample size (Hojat & Xu, 2004), and the degree to which sample results diverge from expectations in the null hypothesis (Vacha-Haase & Thompson, 2004). Kelley and Preacher (2012) summarize these various conceptualizations of effect size and offer a more inclusive definition of effect sizes as a "quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest" (p.140).

Lakens (2013) describes effect sizes as among the most important outcomes to report in empirical studies. Effect sizes are important because they provide evidence of practical significance by representing the magnitude and direction of a relationship or difference, often in standardized metrics which can be understood regardless of the scale used (Kirk, 1996). Standardized effect sizes are particularly useful with abstract measurement indices, like those often found in survey research (e.g., NSSE's Engagement Indicators), because they convert raw differences to a standardized metric that can be compared across studies.

This is not to say that standardized effect sizes are always the most appropriate or useful expression of results. Indeed, when the underlying metric is meaningful in terms of its unit of measurement (enrollments, expenditures, hours, etc.), raw difference effect sizes can be more useful and easier to interpret than a standardized effect size (Lipsey et al., 2012). Too often, higher education research does not generate concrete measurement indices so we rely on standardized effect sizes, which are the focus of this article.

## Criticisms of null-hypothesis significance testing

Criticisms of null-hypothesis significance testing (NHST) are not new (e.g., Cohen, 1994; Ferguson, 2009; Hill & Thompson, 2004; Kirk, 1996; Kline, 2013; Wasserstein & Lazar, 2016). Scholars have long regarded NHST as imperfect for examining data, yet the discussion on the meaning and use of statistical significance continues to this day. Recently, the American Statistical Association (ASA, Wasserstein & Lazar, 2016) published a set of guidelines regarding the use and misuse of *p*-values. These critiques of NHST can be summarized in three main criticisms. The first concerns a misunderstanding of *p*-values. In NHST, the *p*-value gives the mathematical likelihood or probability of obtaining these data or more extreme data (D) given that the null hypothesis ($H_0$) is true—that is, $P(D|H_0)$. However, researchers sometimes misinterpret the *p*-value from statistical tests to mean the probability the null hypothesis is true given that we have observed these data—that is, $P(H_0|D)$ (Cohen, 1994; Kirk, 1996; Kline, 2013; Wasserstein & Lazar, 2016). Unfortunately for researchers $P(D|H_0) \neq P(H_0|D)$ ; nor does obtaining data with a small $P(D|H_0)$ imply that $P(H_0|D)$ is also small (Cohen, 1994; Kirk, 1996). The main criticism here is that NHST does not tell us what we really want to know, whether or not the null hypothesis is true (Ferguson, 2009).

A second criticism is that NHST is very sensitive to sample size. Given a large enough sample, nearly any statistic can be found to be statistically significant. Because sample size is part of the calculation of the standard error, as the number of cases increases the standard error becomes smaller and the test statistic becomes larger, thus making it easier to find statistical significance. As Thompson (1998) quipped, "If we fail to reject, it is only because we've been too lazy to drag in enough participants" (p. 799). This feature is not necessarily a flaw of the hypothesis testing but rather is how the hypothesis test was designed to work.

**NSSE also provides comparative data on these measures from other participating campuses (in aggregate). Such comparisons are valuable to know where to direct institutional improvement efforts.**

This brings us to our third criticism of NHST—statistical significance does not equal practical significance. People often trumpet a small *p*-value (e.g., $p<.001$) as if it indicates a particularly large effect (Kirk, 1996; Lipsey et al., 2012; Wasserstein & Lazar, 2016). Statistical significance evaluates the probability of sample results but it does not tell us whether the effects are substantively important—an issue of greater interest to assessment professionals and policymakers. Statistical significance merely represents statistical rareness, but unlikely events can be completely meaningless or trivial, and conversely, likely events may be quite noteworthy. Unfortunately, *p*-values are confounded by the joint influences of sample results and sample size. Therefore, we use effect sizes to gauge the practical importance of results.

## Types of effect sizes

Effect sizes are generally classified into three broad categories, generally understood as (a) measures of difference, (b) measures of strength of association, and (c) other measures (e.g., Fritz, Morris, & Richler, 2012; Kirk,1996; Rosnow & Rosenthal, 2003; Vacha-Haase & Thompson, 2004). 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 group scores, and include raw differences (e.g., $Mean_1 - Mean_2$), standardized differences (e.g., Cohen's *d*, Hedges' *g*, 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 Pearson's *r*, the popular Pearson product-moment correlation coefficient. This family of measures is concerned with measures of correlation and variance explained and includes such statistics as Pearson's *r*, $r^2$, eta-squared ($\eta^2$), partial

eta square ($\eta_p^2$), and omega-squared ($\omega^2$). The third category often serves as a catchall and includes other measures of effect such as risk estimates like the odds ratio, relative risk, or risk difference.

Results from student engagement comparisons are generally measures of difference, so we focus in this article on two *d*-type effect sizes, Cohen's *d* and Cohen's *h*. Cohen's *d* is used to describe the standardized mean difference between the scores of two groups of independent observations. It is calculated by dividing the mean difference by the pooled standard deviation. While it was Hedges (1982) who first 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 et al., 2012). The formula to compute Cohen's *d* is as follows:

$$d = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{(n_1-1)\,s_1^2 + (n_2-1)\,s_2^2}{n_1 + n_2 - 2}}}$$

where ($\overline{X}_j$) is the sample mean for the $j^{th}$ group, $s_j^2$ is the sample variance for the $j^{th}$ group, and $n_j$ is the sample size for the $j^{th}$ group. The denominator is often referred to as the pooled estimate of standard deviation ($s_{pooled}$) and is the square root of the unbiased estimate of the within-group variance (Kelley & Preacher, 2012).

**Standardized effect sizes are particularly useful with abstract measurement indices, like those often found in survey research (e.g., NSSE's Engagement Indicators), because they convert raw differences to a standardized metric that can be compared across studies. This is not to say that standardized effect sizes are always the most appropriate or useful expression of results.**

Cohen's *h* effect size is the difference between two independent proportions (e.g., the percentage of students who participated in a particular activity such as study abroad or an internship) after each proportion has been transformed using an arcsine transformation. Specifically, it is calculated as follows:

$$h = (2\sin^{-1})\sqrt{P_1} - (2\sin^{-1}\sqrt{P_2})$$

where $P_j$ is the sample proportion for the $j^{th}$ group. 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; Hojat & Xu, 2004; Rosnow & Rosenthal, 2003). This type of transformation is known as a variance stabilizing transformation. Since the variance of a proportion is equal to the proportion multiplied by one minus the proportion divided by the sample size [$VAR\,(p) = \frac{(p)\,(1-p)}{n}$ where $p$ represents the proportion and $n$ represents the sample size], 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. 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 $\pi$, or around -3.14 to 3.14; this is because values of the arcsine function range between $-\pi/2$ and $\pi/2$.

## Interpreting effect sizes

The purpose of reporting effect sizes is for a reader to better judge the importance of the findings. However, in order to understand the importance of results for abstract measurement indices such as the NSSE Engagement Indicators, the effect size must be contextualized against some frame of reference. The most popular frame of reference—a set of benchmarks offered by Cohen (1988, 1992) — is also common in educational research (see, McMillan & Foley, 2011; Peng, Chen, Chiang, & Chiang, 2013 for a review of effect size reporting in major journals). Cohen 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* and *h* values of .2, .5, and .8, and *r* values of .1, .3, and .5, would represent small, medium, and large effects respectively. Yet, Cohen (1988) 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*, even suggesting to researchers to "avoid the use of these conventions, if he can, in favor of exact values provided by theory or experience in the specific area in which he is working" (p. 184). Further complicating the interpretation 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, converting *r* to *d* using the formula provided by Cohen (1988, p. 23), $r = (d / \sqrt{d^2+4})$, we see 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. Nevertheless, Cohen's recommendation has been incorporated into many educational, behavioral, and social science studies.

While discussing interpretations of effect sizes, Cohen (1988) cautioned that when a construct cannot be brought into the laboratory to be studied, which is the case in the vast majority of higher education assessments, extraneous or uncontrollable factors could lead to smaller or more difficult-to-detect effect sizes. In the realm of educational research, Cohen was right. For example, Hill, Bloom, Black, and Lipsey (2008) summarized estimates of achievement effect sizes from studies of K-12 educational interventions and noted that the standardized mean differences (Cohen's *d*) typically ranged from .20 to .30. Similarly, investigating K-12 students' academic performance on standardized reading and mathematics achievement tests, Lipsey et al. (2012) found standardized mean differences as large as .30 to be rare. When investigating school-level performance gaps, Bloom, Hill, Black, and Lipsey (2008) found standardized mean differences between "weak" (i.e., 10th percentile) and "average" (i.e., 50th percentile) schools to be in the .20 to .40 range.

**Statistical significance evaluates the probability of sample results but it does not tell us whether the effects are substantively important—an issue of greater interest to assessment professionals and policymakers.**

Researchers in other social and behavioral sciences have also noted that study effects were often small by Cohen's standards. Ellis (2010a) investigated the average effect size in international business research from 1995 to 2009 and found typically small effect sizes ($r < .10$) by Cohen's standards. Rosnow and Rosenthal (1989, 2003) note that small effect sizes are not that unusual in biomedical research. They illustrate how a seemingly trivial or very small effect can have important real-life consequences. For example, in a study to examine the effects of aspirin on incidence of heart attacks, an effect size of $r = 0.034$ was used to end the study prematurely because it had become clear that aspirin prevents heart attacks and it would have been unethical to continue to give half the participants a placebo. Rosnow and Rosenthal (1989, 2003) argue that this is not to suggest that all small effects are noteworthy; rather, that small effects can have practical consequences in life and death situations. They conclude that in research involving hard-to-change outcomes, such as the incidence of heart attacks, small effects can have profound practical significance.

Few of the effects mentioned above would be described as anything other than small by Cohen's (1988, 1992) standards. What can be taken from these examples is that the interpretation of effect sizes is context dependent. In fact, many scholars (e.g., Cohen, 1988; Hill & Thompson, 2004; Kelley & Preacher, 2012; Kirk, 1996; Thompson, 2001; Vacha-Haase & Thompson, 2004) criticize the use of universally accepted guidelines, like Cohen's benchmarks, for interpreting effect sizes. As Thompson (2001) points out, "if people interpreted effect sizes with the same rigidity that $\alpha = .05$ has been used in statistical testing, we would merely be being stupid in another metric" (p. 82-83).

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 previous studies are

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 research that statistical results be accompanied by an effect size and a "qualitative interpretation" of the effect. These recommendations have been endorsed by journal editors in higher education (e.g., Smart, 2005) and other behavioral and social science disciplines (e.g., Lopéz, Valenzuela, Nussbaum, & Tsai, 2015; Vacha-Haase & Thompson, 2004) who have also called on distinguishing between the statistical and practical significance of a study's findings. Unfortunately, most research in education utilizes Cohen's recommendations of small, medium, and large effects rather than interpreting the effect size within the context of previous findings or research (McMillan & Foley, 2011; Peng et al., 2013). Given the importance of contextualizing an effect within a specific research area, assessment professionals, researchers, and policymakers assessing student engagement need the ability to interpret effect sizes of their results within the context of other student engagement results.

**The purpose of reporting effect sizes is for a reader to better judge the importance of the findings. However, in order to understand the importance of results for abstract measurement indices such as the NSSE Engagement Indicators, the effect size must be contextualized against some frame of reference.**

## Purpose and Research Questions

The purpose of this study is to examine the distribution of effect sizes derived from institutional comparisons from the National Survey of Student Engagement (NSSE) and to make recommendations for their interpretation. 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 interpretations that fit the context of institutional engagement results?

# Method

## Data Source

The NSSE data used in this study were obtained and used with permission from The Indiana University Center for Postsecondary Research. As mentioned previously, NSSE is an annual survey administered to first-year and senior students at baccalaureate degree-granting colleges and universities and is used to assess the extent to which students are exposed to and participate in effective educational practices (McCormick et al., 2013). The analytic sample consisted of 984 U.S. institutions that participated in the 2013 or 2014 administration of NSSE. For institutions that participated both years, we only included the 2014 data. Participating institutions represented a broad cross-section of the national profile of U.S. bachelor's degree-granting institutions (Table 1).

## Measures

Effect sizes for the study were based on comparisons of two primary sets of variables generated from the NSSE questionnaire: Engagement Indicators (EIs) and High-Impact Practices (HIPs). NSSE's 10 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. The psychometric properties of these measures have been described in detail elsewhere (BrckaLorenz & Gonyea, 2014; Miller, Sarraf, Dumford, & Rocconi, 2016).

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

| | | % |
|---|---|---|
| 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 |

**Given the importance of contextualizing an effect within a specific research area, assessment professionals, researchers, and policymakers assessing student engagement need the ability to interpret effect sizes of their results within the context of other student engagement results.**

HIPs encompass several co-curricular educational experiences that have been recognized as "high-impact" due to their positive associations with student learning and development in college (Kuh, 2008; Kuh & O'Donnell, 2013). NSSE asks students if they have participated in six HIPs: *learning community, service-learning, research with a faculty member, internship or field experience, study abroad*, and *culminating senior experience*. We excluded comparisons for internships, study abroad, and culminating senior experiences for first-year students because these opportunities are typically not available until later in the undergraduate years.

## Analysis

To answer the first research question, we generated a dataset by calculating effect sizes for each EI and HIP, separately for first-year and senior students, for comparisons of respondents attending each of the 984 institutions with respondents from all other institutions as a single group. Although institutional users of NSSE are allowed to customize comparison groups, we compared results to students enrolled at all other institutions in order to have a common comparison group for analytic consistency. Results were weighted by sex, enrollment status, and institution size (consistent with NSSE reports delivered to institutions).

To answer the second research question, we considered Cohen's (1988) rationale for observing a small effect (i.e., an effect that is hardly noticeable), a medium effect (i.e., an effect that is observable), and a large effect (i.e., an effect that is plainly evident) and considered ways in which such institutional differences would be observable in the data. To accomplish this, we derived a technique 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 average scores of the students attending those institutions is expected to be small. In a similar way, a *medium* effect would resemble the difference between the average scores of students attending institutions in the upper and lower halves of the distribution (Figure 1b), and a *large* effect would resemble the difference between the average scores of students attending institutions in the top and bottom quartiles

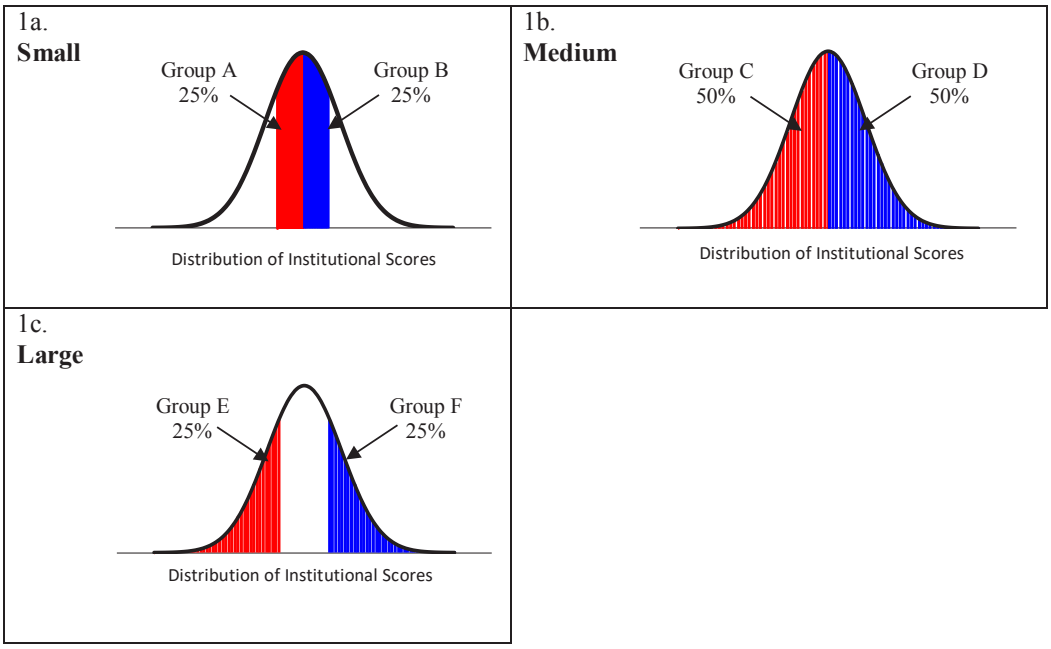


*Figure 1*
Illustration of Three Model Comparison Groups for Determining Empirically-Based Effect Size Thresholds Based on the Distribution of Student Engagement Measures

(Figure 1c). Our analytic approach is similar to a technique used by Bloom et al. (2008) and Konstantopoulos and Hedges (2008) to contextualize effect size estimates for K-12 school achievement in which they estimated differences in achievement for students at "average" (i.e., 50th percentile) and "weak" schools (10th percentile).

The first step in this process was assigning percentile rankings to each of the 984 institution's EI and HIP scores, separately for first-year and senior students. The percentile rankings were based on an institution's precision-weighted score. The precision-weighting process involved adjusting institutional mean scores using Empirical Bayes methods in order to account for lower reliability in institutional means due to small sample sizes and distance from the overall estimate (Hox, 2010). The objective of the precision-weighting adjustment was to avoid over-interpretation of statistical noise in ranking institutions. The precision-weighted means were only used to derive the percentile rankings; unadjusted student-level data were used in the effect size calculations. Once percentile rankings were obtained for institutions' EI and HIP scores, we used these percentile rankings to model effect size comparisons of increasing magnitude (Figure 1). Cohen's *d* and *h* effect sizes were computed according to the formulas presented earlier. For example, to calculate the "small" effect in our proposed scheme, students attending institutions that had percentile ranks between the 50th and 75th percentiles were compared with students attending institutions that had percentile ranks between the 25th and 50th percentiles (Figure 1a). Finally, we calculated confidence intervals for the effect sizes by bootstrapping 1,000 samples for each comparison that was used in each effect size calculation (Kelley & Preacher, 2012).

**These results suggest that new criteria for the interpretation of Cohen's d effect sizes for EIs within the context of NSSE results are necessary.**

Table 2
*Frequency of NSSE Effect Sizes[a] by Cohen's Suggested Ranges[b]*

| | 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[c] | -- | 43% | -- | 38% | -- | 15% | -- | 4% |
| Study Abroad[c] | -- | 40% | -- | 43% | -- | 10% | -- | 7% |
| Culminating Senior Experience[c] | -- | 36% | -- | 36% | -- | 17% | -- | 10% |

[a]Effect sizes were derived from each institution's comparison with the other 983 institutions in the data, separately by class level for each EI and HIP.

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

[c]Effect sizes for Internship, Study Abroad, and Culminating Senior Experience are not calculated for first-year students since these opportunities are typically not available until later in the undergraduate years.

## 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. For most EIs, over 60% of the effect sizes were *trivial* (ES < |.2| in magnitude) and 20% to 30% were *small* (|.2| ≤ ES < |.5|). Only around 1% to 6% of comparisons were within the *medium* range and typically less than 2% met Cohen's criteria of a *large* effect. An exception was Student-Faculty Interaction for seniors, where fewer effect sizes were classified as trivial (41%), and more were classified as medium (16%) and large (4%).

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

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
*Frequency of NSSE EI 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% |

internships and study abroad, and fully 27% of effect sizes for culminating senior experiences were at least medium in magnitude. In contrast, over four-fifths of the institutional comparisons for first-year research with faculty were trivial, and 1% were at least medium in magnitude.

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

**Our study aims to provide assessment professionals, policymakers, researchers and other users of NSSE data a framework to aid in assessing the practical significance of NSSE student 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 our proposed scheme based on the distribution of institutional scores. Table 3 shows the Cohen's *d* effect sizes and confidence intervals for the small, medium, and large model comparisons for first-year and senior students on all 10 EIs. While the effect size estimates 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 a few instances, 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 usually 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 size estimates tended to be lower in nearly every instance.

These results suggest that new criteria for the interpretation of Cohen's *d* effect sizes for EIs within the context of NSSE results are necessary. The consistency of effect size values among the EIs 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. These new reference values were selected after an examination of the effect size values in Table 3, which when rounded to the nearest tenth approximated evenly-spaced intervals between .1 and .5. Table 4 reports the distribution of effect sizes based on the proposed reference values for the Engagement Indicators. As expected from our previous analysis of effect size distribution, the majority of effect sizes were trivial or small. Yet, there is a finer distribution within categories from what we saw in Table 2 based on Cohen's definitions. For the EIs, Table 4 shows that approximately 35% to 40% of all effect sizes were in the trivial range, 40% to 45% were considered small, 10% to 15% were medium, and large effect sizes were relatively rare.

Table 5 shows the Cohen's *h* effect sizes and confidence intervals for the small, medium, and large model comparisons on the six HIPs. Cohen's *h* effect sizes varied more across HIPs and across class year than did the effect size estimates for the EIs. While the effect size estimates for learning communities were generally similar to those of the EIs (.1, .3, and .5), the effect sizes for service-learning, internships, study abroad, and culminating senior experiences were considerably larger and in fact 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 smaller and in a fairly narrow range, with estimates of .06, .17, and .26, respectively. Effect size estimates for research with faculty also varied greatly between class level while estimates for learning community and service-learning were fairly consistent across class level. Average effect sizes for the three first-year HIPs were .11, .31, and .50 for small, medium, and large effects, respectively. Senior estimates for HIP effect sizes were generally larger in magnitude and ranged more. For instance, effect sizes for culminating senior experiences had the largest range, with small, medium, and large effects of .25, .60, and .92, respectively, while learning community effect sizes for seniors had the smallest range, .10, .27, and .43. With the exception of learning community (which typically had lower estimates) and culminating senior experiences (which typically had larger estimates), the other four HIPs for seniors had relatively similar effect size estimates: about .2 for small, between .4 and .5 for medium, and between .6 and .8 for large. Given the variability in Cohen's h effect size estimates both between HIPs and between class levels, it is difficult to provide a set of benchmarks for effect sizes applicable to HIPs in general.

Table 5
*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) |
| Culminating Senior Experience[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 Culminating Senior Experience are not calculated for first-year students since these opportunities are typically not available until later in the undergraduate years.

## Limitations

As with any research, ours is not without its limitations. First, our findings primarily apply to the NSSE Engagement Indicators and High-Impact Practice items. With the exception of the six HIP items, our analysis did not include all the individual items on the NSSE questionnaire. Thus, we urge readers to use caution when applying these recommendations to the individual item estimates in NSSE. Second, Cohen (1988) and others (e.g., Ellis, 2010b; Lakens, 2013; Vacha-Haase & Thompson, 2004) advocate for grounding effects in an area of research; as such we urge caution in applying the study's findings and recommendations on effect sizes to other surveys of undergraduates. Although NSSE is a widely adopted instrument used to assess the student experience, it is only one means by which to measure student engagement, and researchers are encouraged to adopt the study's methods to examine effect sizes in other contexts. Finally, the generalizability of the findings is also limited by the fact that institutions self-selected to participate in NSSE. Although our sample consisted of a wide cross-section of baccalaureate degree-granting institutions (Table 1), it was not necessarily representative of all four-year colleges and universities in the United States. Despite these limitations, we believe this study provides valuable insight to the types of effects that are possible for student engagement results with NSSE data and may guide these professionals in their interpretation of student engagement results.

**Despite these limitations, we believe this study provides valuable insight to the types of effects that are possible for student engagement results with NSSE data and may guide these professionals in their interpretation of student engagement results.**

## Discussion

Knowing whether an institution scored statistically higher than its comparison group on a particular Engagement Indicator (EI) is not particularly helpful to an assessment professional or administrator. At the same time, raw score differences for abstract indices, like NSSE's Engagement Indicators, are difficult to interpret because they lack a meaningful measurement unit. Therefore, in order to communicate the importance of engagement survey results to assessment professionals, policymakers, and other users of NSSE, statistical comparisons need to be translated into a form that facilitates more practical interpretations. While professional organizations (e.g., AERA, APA, ASA) and journal editors (e.g., Smart, 2005; Lopéz et al. 2015) call for researchers to report effect sizes in their studies, researchers infrequently interpret what they mean or compare them to previous effects (Lakens, 2013; McMillan & Foley, 2011;

Peng et al., 2013). Absent a meaningful context grounded in data that are common to the field or area of research, an effect size by itself provides very little other than transforming the difference into standardized units. Interpreting the magnitude or practical significance of an effect size requires it to be compared with other appropriate effects that are relevant to the research study (Kelley & Preacher, 2012; Lipsey et al., 2012; Vacha-Haase & Thompson, 2004). Our study aims to provide assessment professionals, policymakers, researchers and other users of NSSE data a framework to aid in assessing the practical significance of NSSE student engagement results.

**These effect size recommendations are not intended to be definitive judgments on the relative efficacy of NSSE's Engagement Indicators.**

Our findings reinforce Cohen's (1988) caution against the use of universal benchmarks for interpreting effect sizes. Results from our study indicated that Cohen's benchmarks did not adequately fit effect sizes seen in NSSE, especially for the EIs. When examining the distribution of effect sizes within Cohen's benchmarks (Table 2), nearly all effects achieved would be considered trivial or small. Rarely did effect size estimates meet Cohen's thresholds for medium and large, particularly for the EIs. Using our contrived comparisons to mimic effect sizes of increasing magnitude, we found that the EIs could be better summarized using a .1, .3, .5 convention for small, medium, and large effects, respectively. Like Cohen's benchmarks, 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.

The proposed values for EIs may have intuitive and functional appeal for assessment professionals and other users of NSSE data. They are grounded in actual NSSE data, which allows for richer interpretations of the results. Institutions with meaningful differences will more likely find effect sizes of .3 or .5 and can be more confident in interpreting those effects as medium or large effects. Furthermore, although relatively small, one should not simply disregard effect sizes of .1 as trivial. In their review of psychological, educational, and behavioral treatment interventions, Lipsey and Wilson (1993) reached similar conclusions regarding findings with small effect sizes stating, "we cannot arbitrarily dismiss modest values (even 0.10 or 0.20 SDs) as obviously trivial" (p. 1199). Similarly, in their study of school reform, Konstantopoulos and Hedges (2008) remark that an effect of half a standard deviation (i.e., $d$ = .5) should be interpreted as a very large effect in the context of school reform.

A goal of this article is to provide assessment professionals, policy makers, and researchers guidelines for interpreting NSSE student engagement effects sizes. Assessment professionals, in particular, can utilize these results by using effect sizes for guidance on which items to report to stakeholders. They can use our contextualized results and recommendations to identify areas of engagement where an institution is doing comparatively well, and to identify areas in need of improvement. For example, finding a negative, medium in magnitude effect size (such as -.30) in comparison to a group of peer institutions on the Student-Faculty Interaction indicator, an institution might set a goal to improve the quality and frequency of contact between students and faculty. Our findings can aid users in answering what is a meaningful difference, and what effect sizes are typical in this area?

These effect size recommendations are not intended to be definitive judgments on the relative efficacy of NSSE's Engagement Indicators. As Hill et al. (2008, p.176) states, "empirical benchmarks from a research synthesis do not indicate what effects are desirable from a policy standpoint;" instead, they serve to indicate what effects are likely and attainable. Our recommended benchmarks are a general gauge but can provide some guidance as to what magnitude effects are typical with student engagement results and NSSE data in particular.

Our effect size comparisons are most appropriate to serve as a reference for making institution-to-peer comparisons for the EI and HIP items on NSSE. While our analyses focused on comparisons among institutions, intra-institutional comparisons (e.g., comparisons across years, major fields of study, co-curricular involvement) are also often important and interesting to assessment professionals. Although our analyses did not focus on intra-institutional comparisons, our findings may be useful as a starting point when investigating these relationships since our results are grounded in NSSE data. However, we caution readers when making these comparisons that knowledge of the subject matter, and not blind reference to our findings, is warranted. For instance, an assessment professional interested in how often

students use quantitative reasoning skills across academic majors should keep in mind that certain majors emphasize these skills more than others (Rocconi, Lambert, McCormick, & Sarraf, 2013), and as such, should expect larger effect size differences among certain academic majors (e.g., humanities compared with physical sciences). Future research in this area needs to consider these intra-institutional comparisons.

For researchers or users interested in a specific EI, referring to the results in Table 3 would offer more accurate or meaningful information on the estimate of effect size for a particular indicator. Our recommended benchmarks fit better for some EIs than others. For instance, the Discussions with Diverse Others, Quality of Interactions, and Supportive Environment indicators closely follow the new recommended pattern of .1 for small, .3 for medium, and .5 for large. However, some indicators had effects slightly smaller than the recommended cut-off points. For instance, the largest effects for Higher-Order Learning, Reflective and Integrative Learning, and Learning Strategies were between .31 and .41. On the other hand, Student-Faculty Interaction and Collaborative Learning had slightly higher effect size estimates than the recommended benchmark values. Student-Faculty Interaction for seniors particularly stands out as an exception to our general guidelines with estimated effects closer to Cohen's recommendations of small, medium, and large: .2, .5, .8, respectively.

**Our recommended benchmarks are a general gauge but can provide some guidance as to what magnitude effects are typical with student engagement results and NSSE data in particular.**

We were unable to recommend a new set of benchmarks for interpreting the results from HIP comparisons. The effect size estimates among the HIPs and between class years varied so greatly that it was difficult to reduce them into a general recommendation for all HIPs. We encourage researchers and users of NSSE data to examine the effect size estimates in Table 5 to gauge the size or practical importance for a particular high-impact practice.

The effect size estimates we found were consistent with the claims of prior researchers in education and the social and behavioral sciences who found effect sizes rarely as large as Cohen's suggestions and often variable from one context to another (e.g., Bloom et al., 2008; Ellis, 2010a; Hill et al., 2008; Lipsey et al., 2012; Rosnow & Rosenthal, 1989, 2003). One reason the effect size estimates for the EIs were generally smaller in magnitude, compared with most of the HIPs, is because they are more abstract concepts, as opposed to the HIPs which are more concrete educational outcomes. Cohen (1988) cautioned that with more abstract and difficult to measure phenomena, the statistical noise brought on by uncontrollable factors and measurement error can lead to smaller effect sizes. Compared with the EIs, institutions have more direct control over HIPs. Program faculty or other institutional leaders can implement policies that require seniors to complete a culminating thesis or that implement a college-wide initiative with a service-learning component. In addition, HIPs are measured using a single item on the survey while the EIs are a collection of individual items used to create a scale measuring the desired construct.

**The ability to contextualize effect sizes found in NSSE will aid assessment professionals and policymakers in judging the relative importance of student engagement results within the context of the survey and better enable these professionals to make more informed decisions on the relative size and practical value of student engagement results.**

As Ellis (2010b) argues, effect sizes are "meaningless unless they can be contextualized against some frame of reference" (p. 32). Unfortunately, contextualizing the meaning of an effect grounded within the specific research context is not that common in the educational research literature (see McMillan & Foley, 2011; Peng et al., 2013). Our study provides researchers and users of NSSE the ability to contextualize the effects found in their studies against a frame of reference grounded in actual NSSE data. Contextualizing the interpretations of effect sizes not only helps facilitate the interpretation of results but can also aid researchers in building on previous findings. Our study provided new guidelines for considering the size of effects with NSSE's EI and HIP data. We believe the empirical results we have presented provide better guidance to a user of NSSE data than the conventional guidelines provided by Cohen. The ability to contextualize effect sizes found in NSSE will aid assessment professionals and policymakers in judging the relative importance of student engagement results within the context of the survey and better enable these professionals to make more informed decisions on the relative size and practical value of student engagement results.

## References


American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher, 35*(6), 33–40.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, D. C.

Astin, A. W. (1993). *What matters in college? Four critical years revisited*. San Francisco: Jossey-Bass.

Bloom, H. S., Hill, C. J., Black, A. B., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness, 1*(4), 289¬328.

BrckaLorenz, A., & Gonyea, R. M. (2014). *The NSSE update: Analysis and design of ten new engagement indicators*. Bloomington, IN: Indiana University Center for Postsecondary Research.

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. (2010a). Effect sizes and the interpretation of research results in international business. *Journal of International Business Studies, 47*(9), 1581–1588.

Ellis, P. D. (2010b). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. New York: Cambridge University Press.

Fisher, R. A. (1925). *Statistical methods for research workers*. London: Oliver & Boyd.

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

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology, 141*(1), 2–18.

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* (2nd ed.). New York: Routledge.

Hedges, L. V. (1982). Estimation of effect size from a series of independent observations. *Psychological Bulletin, 92*(2), 490-499.

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.

Hojat, M., & Xu, G. (2004). A visitor's guide to effect size. Advances in Health Sciences Education, 9(3), 241-249.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York: Routledge.

Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement, 6*(2), 227–24.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17*(2), 137¬153.

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 ed.). Washington, DC: American Psychological Association.

Konstantopoulos, S., & Hedges, L. V. (2008). How large an effect can we expect from school reform? *Teachers College Record, 110*(8), 1613–1640.

Kuh, G. D. (2008). *High-impact educational practices: What they are, who has access to them, and why they matter*. Association of American Colleges and Universities.

Kuh, G. D., & O'Donnell, K. (2013). *Ensuring quality & taking high-impact practices to scale*. Washington, DC: Association of American Colleges and Universities.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-test and ANOVAs. *Frontiers in Psychology, 4*, 1–12.

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.

Lipsey, M., & Wilson, D. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*(12), 1181–1209.

Lopéz, X., Valenzuela, J., Nussbaum, M., & Tsai, C. (2015). Some recommendations for the reporting of quantitative studies. *Computers & Education, 91*, 106-110.

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 (Vol. 28, pp. 47–92). Dordrecht, The Netherlands: Springer.

McMillan, J. H., & Foley, J. (2011). Reporting and discussion effect size: Still the road less traveled? Practical Assessment, *Research & Evaluation, 16*(14), 1–12.

Miller, A. L., Sarraf, S. A., Dumford, A. D., & Rocconi, L. M. (2016). *Construct validity of NSSE Engagement Indicators*. Bloomington, IN: Center for Postsecondary Research.

National Survey of Student Engagement. (2015). *Lessons from the field–Volume 3: Using data to catalyze change on campus*. Bloomington, IN: Center for Postsecondary Research, Indiana University School of Education.

National Survey of Student Engagement. (2017). *Lessons from the field–Volume 4: Digging deeper to focus and extend data use*. Bloomington, IN: Center for Postsecondary Research, Indiana University School of Education.

Pace, C. R. (1979). *Measuring outcomes of college: Fifty years of findings and recommendations for the future*. San Francisco: Jossey-Bass.

Pascarella, E. T., & Terenzini, P. T. (1991). *How college affects students: Findings and insights from twenty years of research*. San Francisco: Jossey-Bass.

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

Pearson, K. (1900). On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London, 195*, 1–47.

Peng, C. J., Chen, L., Chiang, H, & Chiang, Y. (2013). The impact of APA and AERA guidelines on effect size reporting. *Educational Psychological Review, 25*(2), 157–209.

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), Article 10. DOI: http://dx.doi.org/10.5038/1936-4660.6.2.10

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*(10), 1276–1284.

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

Smart, J. C. (2005). Attributes of exemplary research manuscripts employing quantitative analyses. *Research in Higher Education, 46*(4), 461–477.

Springer, R. (2006). Using effect size in NSSE survey reporting. *Research & Practice in Assessment, 1*, 18–22.

Thompson, B. (1998). In praise of brilliance: Where the praise really belongs. *American Psychologist, 53*(7), 799–80.

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education, 70*(1), 80–93.

Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology, 51*(4), 473–481.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133.

Wilkinson, L., & American Psychological Association Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594–604.