



# Contextualizing NSSE Effect Sizes: Empirical Analysis and Interpretation of Benchmark Comparisons

NSSE staff are frequently asked to help interpret effect sizes. “Is .3 a small effect size?” “Is .5 a really large effect size?” An effect size (ES) is any measure of the strength of a relationship between two variables. In practice ES statistics are used to assess comparisons involving correlations, percentages, mean differences, probabilities, and so on. In cases where large sample sizes make it more likely that a difference – even a small one – will be *statistically* significant, ES statistics are often thought of as a measure of *practical* significance because they indicate the relative magnitude of the difference. Thus they are valuable in comparing abstract measurement indices such as the NSSE benchmarks which are computed on a 0 to 100 scale from sets of individual items using various response sets.

NSSE’s comparison reports use Cohen’s *d*, the standardized difference between the institution’s mean and the comparison group’s mean, calculated by dividing the mean difference by the pooled standard deviation. Thus, ES is discussed in this guide solely in terms of the Cohen’s *d* statistic. In his classic book, Cohen (1988) reluctantly defined ES as "small, *d* = .2," "medium, *d* = .5," and "large, *d* = .8," preferring to be intentionally vague about precise cut points and decision rules. Cohen also said that "there is a certain risk inherent in offering conventional operational definitions for those terms used in power analysis in as diverse a field of inquiry as behavioral science" (p. 25) and urged researchers to interpret ES based on the *context of the data*. Nevertheless, Cohen’s definition of small, medium, and large has been widely accepted and incorporated into many social science studies that report ES.

In the following sections, Cohen’s cut points and an empirically derived set of cut points are used to examine the distribution of effect sizes from the *NSSE 2007 Benchmark Comparisons* reports delivered to participating institutions (N=587). Following the analysis, we offer recommendations for interpreting the effect sizes of NSSE benchmark comparisons.

## Frequency of Different Effect Sizes Based on Cohen’s General Definition

Table 1 shows the percentages of 2007 institutions that had effect sizes within Cohen’s cut point ranges on each of the five NSSE benchmarks for first-year (FY) and senior (SR) students. Effect sizes in this table are drawn from the individual institutions’ comparisons with the entire NSSE 2007 cohort. The table shows that the vast majority of effect sizes on benchmark reports were either *trivial* (less than .20 in magnitude) or *small* (.20 to .49 in magnitude). Very few institutions found *medium* or *large* effect sizes using Cohen’s rule-of-thumb criteria.

**Table 1**  
Distribution of NSSE Effect Sizes by Cohen’s General Definition

	Effect Size Range <sup>a</sup>							
	Trivial (< .20)		Small (.20 to .49)		Medium (.50 to .79)		Large (.80 or greater)	
	FY	SR	FY	SR	FY	SR	FY	SR
Level of Academic Challenge	50%	62%	42%	30%	7%	7%	1%	1%
Active & Collaborative Learning	54%	56%	37%	36%	7%	7%	2%	1%
Student-Faculty Interaction	60%	48%	34%	38%	6%	11%	1%	3%
Enriching Educational Experiences	52%	40%	40%	37%	7%	15%	1%	8%
Supportive Campus Environment	50%	46%	43%	44%	7%	9%	1%	1%

<sup>a</sup> Effect sizes were taken only from those NSSE 2007 institutions that selected comparisons with the entire 2007 U.S. NSSE cohort (n=519). Because effects sizes can be both positive and negative, absolute values were used for the ranges.

## Effect Size Interpretation Based on NSSE Data

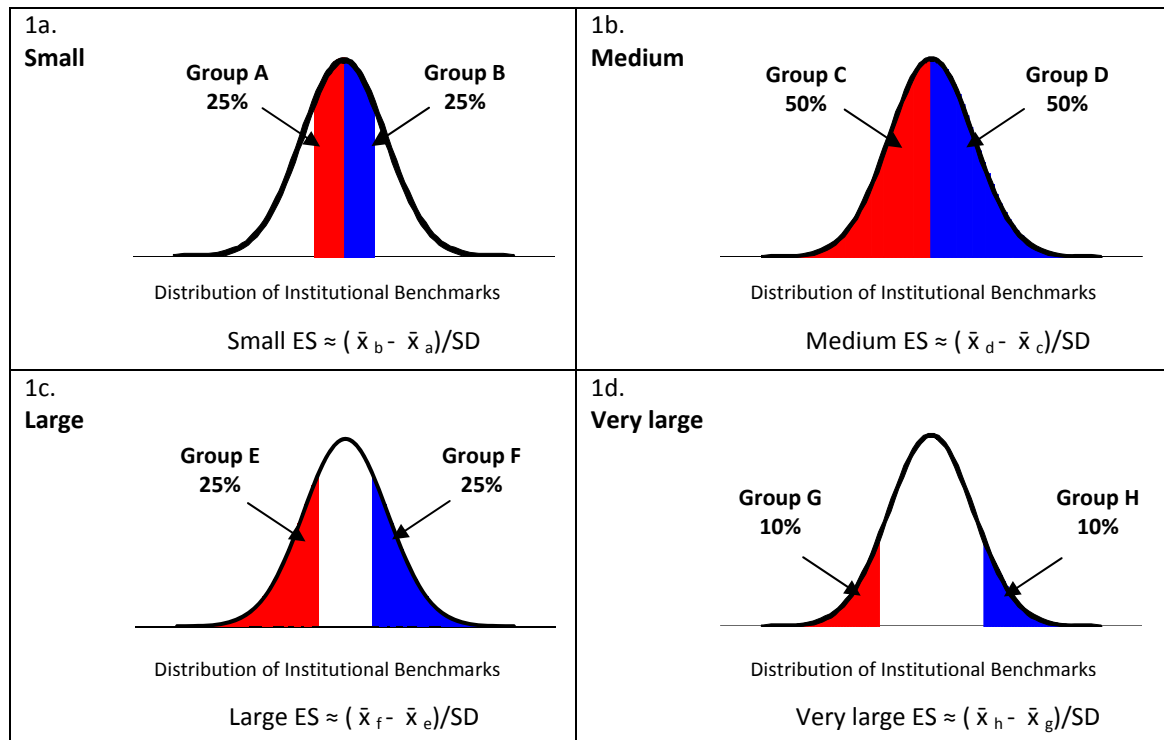
Cohen described *small* effects as those that are hardly visible, *medium* effects as observable and noticeable to the eye of the beholder, and *large* effects as plainly evident or obvious. With respect to this rationale, NSSE staff considered ways in which benchmark differences would be observable in the data, and proposed a scheme to interpret effect sizes based on the distribution of actual benchmark scores. To examine this alternative scheme, NSSE analysts assigned percentile rankings to institutions' 2007 benchmark scores and used them to model comparisons that would resemble effect sizes of increasing magnitude (illustrated in Figures 1a – 1d).

To explain, suppose that a *small* ES would resemble the difference between the benchmark scores of students attending institutions in the third quartile (i.e., between the 50<sup>th</sup> and 75<sup>th</sup> percentiles) and those attending institutions in the second quartile (i.e., between the 25<sup>th</sup> and 50<sup>th</sup> percentiles). 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 the same way, a *medium* ES (Figure 1b) would look like the difference between the benchmark scores of students attending institutions in the upper half (Group D) and those attending institutions in the lower half (Group C) of the distribution. A *large* ES (Figure 1c) would resemble the difference between students attending institutions scoring in the top quartile (Group F) and those attending institutions scoring in the bottom quartile (Group E) of the distribution. And finally, a *very large* ES (Figure 1d) would be like the difference between students attending institutions scoring in the lowest 10% (Group H) and highest 10% (Group G) of the distribution.

### Figures 1a- 1d

Illustration of Four Model Comparison Groups for Determining Empirically-Based Effect Size Thresholds Based on the Distribution of NSSE Benchmarks



Note:  $\bar{x}_a$  through  $\bar{x}_h$  are the mean benchmark scores of the students attending institutions within groups A through H.  
SD=Standard deviation (pooled).

Below each figure is a formula for calculating the Cohen's *d* effect size for the particular comparison being modeled in the illustration. For example, the formula under Figure 1a ( $\text{Small ES} \approx (\bar{x}_b - \bar{x}_a)/\text{SD}$ ) conveys that a small ES is approximately the standardized difference between the mean benchmark score of students attending institutions in Group B ( $\bar{x}_b$ ) and the mean benchmark score of students attending institutions in Group A ( $\bar{x}_a$ ).

Table 2 shows the effect sizes for these *small*, *medium*, *large*, and *very large* model comparisons for first-year students and seniors on all five NSSE benchmarks from 2007. While the effect sizes in Table 2 vary somewhat between benchmarks and between student class levels, the ranges within the *small*, *medium*, *large*, and *very large* categories are consistent and, with the exception of Enriching Educational Experiences for seniors, do not overlap. That is, the maximum *small* ES is lower than the minimum *medium* ES, the maximum *medium* ES is lower than the minimum *large* ES, and so on.

**Table 2**  
Effect Sizes from the NSSE Benchmark Percentile Group Comparisons<sup>a</sup>

	First-year				Seniors			
	Small <sup>b</sup>	Medium	Large	Very large	Small	Medium	Large	Very large
Level of Academic Challenge	0.18	0.42	0.60	0.75	0.12	0.32	0.51	0.74
Active & Collaborative Learning	0.13	0.37	0.56	0.70	0.11	0.35	0.52	0.69
Student-Faculty Interaction	0.13	0.34	0.49	0.63	0.14	0.39	0.63	0.87
Enriching Educational Experiences	0.16	0.38	0.56	0.77	0.24	0.54	0.87	0.99
Supportive Campus Environment	0.18	0.41	0.57	0.72	0.16	0.45	0.61	0.76
<i>Minimum</i>	<i>0.13</i>	<i>0.34</i>	<i>0.49</i>	<i>0.63</i>	<i>0.11</i>	<i>0.32</i>	<i>0.51</i>	<i>0.69</i>
<i>Maximum</i>	<i>0.18</i>	<i>0.42</i>	<i>0.60</i>	<i>0.77</i>	<i>0.24</i>	<i>0.54</i>	<i>0.87</i>	<i>0.99</i>

<sup>a</sup> Only U.S. NSSE 2007 institutions were included in this analysis (n=587). Precision-weighted means were used to determine the percentile ranking of each institution's benchmark, separately for first-year and senior students.

<sup>b</sup> *Small* = ES difference between the scores of students attending institutions in the third quartile and those of students attending institutions in the second quartile on a particular benchmark. *Medium* = ES difference between the scores of students attending institutions in the upper half and those of students attending institutions in the lower half on a particular benchmark. *Large* = ES difference between the scores of students attending institutions in the top quartile and those of students attending institutions in the bottom quartile on a particular benchmark. *Very large* = ES difference between the scores of students attending institutions in the top 10% and those of students attending institutions in the bottom 10% on a particular benchmark.

Tables 1 and 2 suggest that a slightly finer grained approach to effect size interpretation than Cohen's is appropriate for NSSE benchmark comparisons. The consistency of ES values in Table 2 makes it possible to recommend new criteria for the interpretation of effect sizes in benchmark comparisons.

Therefore, Table 3 proposes a new set of reference values for interpreting effect sizes, based on the results in Table 2. 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 ES. These new reference values were selected after an examination of the minimum values in Table 2, which when rounded to the nearest tenth approximated evenly-spaced intervals between .1 and .7. The simplicity of the proposed values (.1, .3, .5, and .7) may have intuitive and functional appeal for users of NSSE data.

**Table 3**

Proposed Reference Values for the Interpretation of Effect Sizes from NSSE Benchmark Comparisons<sup>a</sup>

	Effect size
Small	.1
Medium	.3
Large	.5
Very large	.7

<sup>a</sup> These values were based on NSSE benchmark distributions and are recommended for NSSE benchmark comparisons, not for individual item mean comparisons. Values are to be viewed as coarse thresholds, not as precise cut-points.

Table 4 reports the distribution of NSSE effect sizes based on these proposed reference values. As expected from our previous look at Table 1, the majority of effect sizes were *trivial*, *small*, and *medium*. Yet, this is a finer distribution within categories from what we saw in Table 1 based on Cohen’s definitions. In Table 4 approximately one-quarter to one-third of all effect sizes appear to be in the *trivial* range, more than 40% are considered *small*, and the new *medium* range captures about a 20 to 25% of all effect sizes. *Large* and *very large* effect sizes are relatively rare.

**Table 4**

Distribution of NSSE 2007 Effect Sizes by the Proposed Reference Values

Benchmark	Effect Size Range <sup>a</sup>									
	Trivial (0 to .09)		Small (.10 to .29)		Medium (.30 to .49)		Large (.50 to .69)		Very Large (.70 or more)	
	FY	SR	FY	SR	FY	SR	FY	SR	FY	SR
Level of Academic Challenge	27%	34%	43%	44%	22%	14%	6%	5%	2%	3%
Active & Collaborative Learning	29%	29%	44%	46%	18%	17%	6%	6%	3%	2%
Student-Faculty Interaction	34%	25%	45%	40%	15%	20%	5%	9%	2%	5%
Enriching Educational Experiences	25%	21%	46%	32%	22%	24%	5%	11%	2%	11%
Supportive Campus Environment	27%	23%	41%	42%	24%	25%	7%	7%	1%	3%

<sup>a</sup> Effect sizes were taken only from those institutions that selected comparisons with all 2007 U.S. institutions (n=519). Because effects sizes are both positive and negative, absolute values were used for the ranges.

### Case Study – “Sample University”

The following provides an example of how information provided in this guide can be applied to real results.

Sample University (SU) endeavors to be one of the most engaging institutions in the US, with a challenging and enriching academic experience, active and collaborative students, an open and helpful faculty, and the most supportive infrastructure possible for student learning. After work over several years on several initiatives, the provost asked the director of institutional research to give a progress report based on the latest NSSE results.

Table 5 shows the five benchmark scores for seniors attending Sample University alongside scores for the selected comparison group. The third column lists the effect sizes for the mean comparisons. Of course SU is pleased with these results. Indeed, three of the five benchmarks are substantially positive and affirming of their goals. The director of institutional research noticed that the Level of Academic Challenge at SU is in fact quite strong, with a “very large” effect at .72. Active and Collaborative Learning and Supportive Campus Environment are also well above average with “medium” effects of .44

and .30 respectively. The effect size for Enriching Educational Experiences is also on the positive side, but perhaps “small” in magnitude at .23. The only benchmark showing perhaps no meaningful or practical difference is Student-Faculty Interaction at .08.

**Table 5**  
Sample University Benchmark Comparisons<sup>a</sup> (Seniors)

<i>Benchmark</i>	<i>Benchmark Scores</i>		
	Sample University	Selected comparison group	<i>Effect size</i>
Level of Academic Challenge	65.8	55.6	.72 <i>Very large</i>
Active and Collaborative Learning	57.7	50.1	.44 <i>Medium</i>
Student-Faculty Interaction	42.8	41.2	.08 <i>Trivial</i>
Enriching Educational Experiences	44.0	39.8	.23 <i>Small</i>
Supportive Campus Environment	62.7	56.9	.30 <i>Medium</i>

<sup>a</sup> These results are taken from an actual institution, yet we acknowledge that they are unusually positive. They were intentionally selected for the illustrative purposes of this manuscript.

Although the recommended definitions in the ES chart in Table 3 are useful in interpreting the comparison results, more actionable observations often exist at the item level. Item frequencies can make benchmark scores and effect sizes more tangible and observable. For example, Table 6 reports Sample University’s frequencies for the individual items corresponding to the benchmark scores in Table 5. The response options for all the items were collapsed for quick review and interpretation. Both SU and the selected comparison group percentages are given, with the percentage differences listed in the right hand column. With the exception of items associated with Student-Faculty Interaction, nearly all show positive differences when compared with the selected comparison group. A series of small differences can accumulate into appreciable effect sizes when combined to form the benchmark score.

Among the Level of Academic Challenge items, several large percentage differences stand out for Sample University. For example, 36% more SU students said they read 10 or more assigned books and 27% more wrote at least four mid-length papers. SU seniors also reported that their coursework emphasized substantially more analysis, synthesis, evaluation, and application. These differences in the individual item responses account for the *very large* effect size of .72 on this benchmark.

The *medium* Active and Collaborative Learning effect size of .44 is also evident in the individual item frequencies. For example, compared to seniors attending the selected comparison group institutions, 21% more SU seniors contributed to class discussions frequently (often or very often), and 17% more made class presentations frequently. Likewise, the *medium* effect on the Supportive Campus Environment benchmark is evident in the mostly positive response differences on the individual items, ranging up to 14%.

The *small* magnitude of the Enriching Educational Experiences benchmark is due to mixed results among the items, with those showing positive differences for SU (such as foreign language coursework, internships, and co-curricular activities) to some extent offset by those showing negative differences for SU (such as independent studies and culminating senior experiences). Other items show only modest differences. Still, the net result is a positive effect size of .23.

**Table 6**  
**Sample University Item Frequencies by Benchmark (Seniors)**

<i>Item #</i>	<i>Percent of students who...</i>	Sample University	Selected Comparison Group	Difference
<b>Level of Academic Challenge (ES=.72)</b>				
10a.	Said the institution emphasizes studying and academic work <sup>3</sup>	82%	78%	4%
1r.	Worked harder than you expected to meet an instructor's expectations <sup>1</sup>	66%	57%	10%
2b.	Said courses emphasized analyzing ideas, experiences, or theories <sup>3</sup>	95%	84%	11%
2c.	Said courses emphasized synthesizing ideas into new complex relationships <sup>3</sup>	90%	74%	16%
2d.	Said courses emphasized making judgments about the value of information <sup>3</sup>	86%	71%	15%
2e.	Said courses emphasized applying theories or concepts to new situations <sup>3</sup>	95%	79%	16%
3a.	Read more than 10 assigned books or book-length packs of readings	68%	32%	36%
3c.	Wrote at least one paper or report of 20 pages or more	62%	49%	12%
3d.	Wrote more than 4 papers or reports between 5 and 19 pages	72%	46%	27%
3e.	Wrote more than 10 papers or reports of fewer than 5 pages	37%	31%	6%
9a.	Spent more than 10 hours/week preparing for class (studying, etc.)	69%	55%	14%
<b>Active and Collaborative Learning (ES=.44)</b>				
1a.	Asked questions/contributed to class discussions <sup>1</sup>	90%	69%	21%
1b.	Made a class presentation <sup>1</sup>	76%	59%	17%
1g.	Worked with other students on projects during class <sup>1</sup>	45%	47%	-2%
1h.	Worked with classmates outside of class to prepare class assignments <sup>1</sup>	69%	58%	11%
1j.	Tutored or taught other students (paid or voluntary) <sup>1</sup>	30%	22%	8%
1k.	Did a community-based project as part of a regular course <sup>1</sup>	27%	17%	10%
1t.	Discussed ideas from readings or classes with others outside of class <sup>1</sup>	64%	62%	1%
<b>Student-Faculty Interaction (ES=.08)</b>				
1n.	Discussed grades or assignments with an instructor <sup>1</sup>	57%	58%	-1%
1o.	Talked about career plans with a faculty member or advisor <sup>1</sup>	47%	40%	6%
1p.	Discussed ideas from classes with faculty outside of class <sup>1</sup>	27%	27%	0%
1q.	Received prompt written or oral feedback from faculty <sup>1</sup>	64%	62%	1%
1s.	Worked with faculty members on activities other than coursework <sup>1</sup>	22%	21%	1%
7d.	Worked on a research project with a faculty member outside of class	19%	19%	0%
<b>Enriching Educational Experiences (ES=.23)</b>				
10c.	Said the institution substantially encourages contacts among diverse peers <sup>3</sup>	49%	46%	3%
1l.	Used an electronic medium to discuss or complete an assignment <sup>1</sup>	59%	60%	-1%
1u.	Had serious conversations w/ students of another race or ethnicity <sup>1</sup>	53%	53%	0%
1v.	Had serious conversations w/ students of other relig./politics/values <sup>1</sup>	62%	55%	7%
7a.	Did a practicum, internship, field exp., clinical assgmt	65%	53%	12%
7b.	Participated in community service or volunteer work	68%	59%	9%
7c.	Participated in a learning community	27%	25%	1%
7e.	Completed foreign language coursework	61%	41%	20%
7f.	Completed a study abroad program	23%	14%	9%
7g.	Participated in an independent study or self-designed major	11%	18%	-6%
7h.	Completed a culminating senior experience	23%	32%	-9%
9d.	Spent more than 5 hours/week participating in co-curricular activities	37%	24%	13%
<b>Supportive Campus Environment (ES=.30)</b>				
10b.	Said the institution provides substantial support for academic success <sup>3</sup>	82%	68%	14%
10d.	Said the institution substantially helps students cope w/ non-acad. matters <sup>3</sup>	30%	24%	6%
10e.	Said the institution provides substantial support for students' social needs <sup>3</sup>	40%	34%	6%
8a.	Positively rated their relationships with other students <sup>2</sup>	77%	82%	-4%
8b.	Positively rated their relationships with faculty members <sup>2</sup>	87%	78%	9%
8c.	Positively rated their relationships with admin. personnel and offices <sup>2</sup>	64%	53%	11%

<sup>1</sup> Combination of students responding 'very often' or 'often'

<sup>2</sup> Rated at least 5 on a 7-point scale

<sup>3</sup> Combination of students responding 'very much' or 'quite a bit'

Finally, Student-Faculty Interaction shows a *trivial* effect for SU in comparison with the selected comparison group, which is plainly evident by the meager percentage differences between the two groups.

Taken together, the differences between Sample University and the comparison group in these item frequencies provide a rich explanation for the effect sizes seen in Table 5. Observations like this can help administrators and policy makers cultivate specific action plans to improve the undergraduate experience.

## Conclusion

The purpose of this study was to analyze effect sizes in the context of actual NSSE data and to guide the interpretation of the effect sizes on NSSE's *Benchmark Comparisons* reports. These analyses informed the development of a new set of reference values for interpreting the benchmark effect sizes.

As a practical matter for NSSE users, at least four approaches can be taken with regard to effect sizes.

1. First, it's not unreasonable to continue using Cohen's purposefully vague definition. The new reference values offered in Table 3 only deviate from Cohen in the lower values. Some may be convinced that small effect sizes are unworthy of further examination, and thus should continue to look for values around .5 and greater.
2. Second, for those willing to consider the new reference values proposed in Table 3, the thresholds of .1, .3, .5, and .7 could have appeal for their simplicity and functionality. They are grounded in actual NSSE findings and may allow for richer interpretations of NSSE results.
3. Third, it's also possible to ignore the new reference values and to examine the results in Table 2 for a more nuanced interpretation of a particular ES. Table 2 reveals a different pattern of effect sizes for each benchmark, and also that these differ between first-year students and seniors. What's more, effect sizes for the Enriching Educational Experiences benchmark for seniors tend to be larger in magnitude than for other benchmarks.
4. Finally, the guide also recommends an examination of individual item frequencies in combination with ES 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 ES. For example, consider two institutions with the same ES on a particular benchmark. The first may have large percentage differences on just a few of the benchmark 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 benchmark scores in relation to their selected comparison groups.

## References

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



Indiana University Center for Postsecondary Research  
1900 East Tenth Street, Suite 419  
Bloomington, IN 47406-7512

Phone: 812-856-5824  
Fax: 812-856-5150  
E-mail: [nsse@indiana.edu](mailto:nsse@indiana.edu)  
Web: [www.nsse.iub.edu](http://www.nsse.iub.edu)