Using Effect Size in NSSE Survey Reporting

Robert Springer
Elon University

Abstract

The National Survey of Student Engagement (NSSE) provides participating schools an Institutional Report that includes (among many documents) mean comparisons, frequency distributions, and student respondent data as part of its standard reporting package. Sifting through all this data can leave even experienced researchers wondering where to start and what to report. For example, how meaningful is it to report frequency percentages or statistically significant differences between your school and other NSSE schools? Fortunately, NSSE also provides an effect size (Cohen's *d*) or practical significance indicator that can help bring context to the results. In addition to its value in conveying NSSE results, using effect sizes in survey research helps to easily identify areas/items of praise as well as areas/items for improvements.

Introduction

Perhaps one of the most overlooked and more useful statistics is effect size. While statistical tests of significance indicate the likelihood that results would differ by chance (and are depend upon sample size), effect size measurements tell us the relative importance or magnitude of the treatment. As a result of the inability of statistical significance to indicate importance or practical significance (Kirk, 1996; Thompson, 1999; Valentine & Cooper, 2003), there is an ongoing debate as to the practical usefulness of statistical significance tests (Hunter 1997; Kirk, 1996;Thompson, 1999), particularly statistical significance tests used as a sole indicator of the meaningfulness of results.

In essence, effect sizes are practical significance/importance indicators (Kirk, 1996; Vacha-Haase & Nilsson, 1998; Valentine & Cooper, 2003). In a time where collecting large sample sizes has become relatively easy and affordable (i.e. web-based surveys), it is important to distinguish between statistical significance and practical significance. Consider the following example. Elon University had 950 first-year and senior students participate in 2005 NSSE. Of the 170 items on the NSSE (85 questions for both freshmen and seniors), 114 are statistically significant at the $p < 0.001$ level. So, which of these 114 items should be presented to stakeholders? While mapping survey items to institutional or department purpose is always good practice and would help to identify specific questions for reporting, what about other important items (perhaps equally important to another department/program) that might be slipping through the analysis? Short of presenting the entire NSSE results, which in all likelihood will never be read, what items can be identified that indicate meaningful or practical differences? As previously stated, the effect size indicates the relative importance or magnitude of the difference in scores between a treatment and a non-treatment (control) group. In multi-school surveys, it is reasonable to view your institution as the treatment group and other institutions as the non-treatment group. One reason an institution might participate in a national survey is to compare its results to that of other colleges – to see how they are doing by comparison. One way to evaluate the comparison is to use an effect size to indicate meaningful differences between colleges on particular items/areas?

While there are various types of effect size statistics (e.g., ω^2, adjusted R^2, Hedge's g, Fisher's Z, Glass's Δ, η^2), Cohen's *d* will be the focus of this paper, since it is supplied by NSSE . Cohen's *d* as a reported effect size is becoming very popular (Thalheimer & Cook, 2002). As a result, more and more research is including Cohen's *d* which allows for easier comparisons of the magnitude of treatments across experiments (Thalheimer & Cook, 2002). Besides the advantage of its popularity, this effect size also has theadvantage of allowing comparisons to known benchmarks established by Cohen. He describes a d-value of 0.20 as small, 0.50 as medium (moderate), and 0.80 as large. Adding some perspective to these effect sizes, he states that a moderate effect size is "visible to the naked eye" (Cohen, 1988, p.26).

Purpose

This paper is intended as a best practices presentation by demonstrating the use of effect sizes to assist in reporting. This statistic can quickly identify items of practical significance, which adds to interpretation of results.

NSSE and Cohen's d in Reporting

The National Survey of Student Engagement provides Cohen's *d* in their standard Institutional Report under the Means Comparison results. Thus, any institution that participates in the NSSE will have this statistic provided as part of the standard reporting documentation (in paper and electronic form). Of particular interest is the electronic spreadsheet of the NSSE results. This allows for easy sorting of items by Cohen's *d*.

Effect sizes can be negative. For this to happen, the treatment group is performing at a lesser level than the control group. However, the negative sign could be function of scale direction rather than a perceived lack of performance. For example, in the NSSE question about coming to class unprepared (2005 NSSE item 1f), a negative sign is preferable - meaning that fewer students are coming to class unprepared.

While effect sizes may change from one survey administration to the next, they tend to remain fairly stable. In other words, if the institution has not changed what it is doing with respect to certain items, effect sizes, in all likelihood, will not change from one level to another (small, moderate, or large).

Effect sizes are not new to statistics. Effect sizes can be traced back to at least 1901 with the work of Karl Pearson (Kirk, 1996). Yet as such, reporting effect sizes to stakeholders may not be desirable. Doing so might be confusing and could easily lead to dismissal of the report. Reporting figures that are more widely understood such as percent positive frequencies is more advisable. Stakeholders will understand percent positive frequencies. Elon University uses a percent positive frequency (for example, the number of students that select Very Often or Often for a series of questions). The percent positive scale used in Elon's reporting is believed to be fair and it appears to make sense to various stakeholders.

If survey results are to be used to help make improvements at an institution, it is good practice to identify items where the school does well (areas for praise and celebration) and items where it does not do well (areas for improvement) as compared to other schools. For Elon, items of a practical significance are those that approach or exceed a moderate effect size level ($d > 0.40$). A Cohen's *d* of at least 0.40 is approximately two-thirds the distance between small and moderate levels. As a result, *d* values of at least 0.40 are considered approaching a moderate level.

How do effect size and percent positive frequencies relate to each other? Consider the following example. The criterion of an effect size of 0.40 or higher to report items is applied. One item where first-year student responses met that criterion was item 1h – *worked with classmates outside of class to prepare class assignments*. Its effect size is 0.41, it is statistically significant at the $p<.001$ level, and the percent positive frequency is 63% versus 43% for all NSSE schools. To say that Elon is noticeably different compared to other colleges with respect to this item would probably make sense to the lay person, because they can see the large gap (a 20-point difference) in the percent positive frequencies. In addition, the effect size supports that statement. As a further example, another item where first-year student responses resulted in an area targeted for improvement is item 5, *the extent your examinations during the current school year challenged you to do your best work.* Its effect size was 0.03 (very small), it is not statistically significant, and the percent positive frequency is 54% versus 52% for all NSSE schools. The differences in percent positive frequencies, as well as the small effect size, indicate little if any practical difference exists between the two comparisons. Elon wants academic challenge and rigor to be a hallmark for distinction. Given these results, its first-year students appear to be no more challenged than other schools first year students.

How did we actually use the effect size for reporting purposes? A supplemental two-page report interpreting the NSSE results is sent to senior staff and then to the faculty. The first page of that report provides basic information about NSSE and describes the three tables on the second page. In addition, a short paragraph explains that effect sizes are used to select the items for inclusion into the tables. Since Elon participates in the NSSE each year and in order to address stakeholders possible concerns about effect sizes shifting from year-to-year, we selected items that were extremely consistent with respect to reported effect sizes for a five to six year period depending upon when the item was introduced (that being an effect size equal to or greater than 0.40, or near zero). Table 1 indicates items for first year students that have effect sizes of at least a 0.40 (i.e., high performing items). Table 2 indicates items for senior students that have effect sizes of at least 0.40 (i.e., high performing items). Table 3 indicates items whose effect sizes are at or near zero (low performing items). Each table provides the survey questions and the percent positive frequencies for Elon and all other NSSE schools – effect sizes are not presented (effect sizes are included in the appendices as a point of reference for the reader). The items presented for improvement represent areas that

are important for Elon to achieve in its strategic plan. In general, this two-page report is very effective in convening what Elon does very well (based upon effect size) and what it needs to do better (based upon effect size and areas deemed important by the institution).

Discussion

We hope that readers will see the adaptability in using effect sizes to assist in reporting. For example, while a Cohen's *d* of 0.50 is considered moderate and 0.20 is considered small, we selected effect sizes of 0.40 or higher and those that were near zero. For items where Elon performed well, we simply selected items with effect sizes of 0.40 or higher. Items selected for improvement were also items that are identified as important to Elon. Interesting was the fact that the high performing items tended to confirm institutional belief. This also added support for acceptance of the low performing items.

While not all reporting of survey data should or can be reported using effect sizes, it should be obvious that having NSSE supply effect sizes as part of its standard reporting packages enables a researcher to quickly sort and analyze practical differences between itself and comparison groups. Creating a two-page report is acceptable and desirable at Elon - this may not be the case at other institutions.

National surveys that do not provide an effect size or the statistics necessary to calculate one, would be aiding institutions by adopting such standards in their reports. This would allow true peer/aspirant comparisons on a number of dimensions from students and faculty.

Finally, effect size has much broader implications that just survey data. Many publishers are requesting effect size(s) with interpretation(s) from researchers. As the popularity of reporting effect sizes continues to grow, researchers should take caution intheir interpretations of effects sizes being reported as small, moderate, or large. In other words, let the context of the research help establish typical effect sizes and, therefore, what is worth reporting.

References

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

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, *8,* 3-7.

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

Thalheimer, W. & Cook, S. (2002, August). *How to calculate effect sizes from published research articles: A simplified methodology.* Retrieved November 1, 2005 from http://work-learning.com/effect_sizes.htm.

Thompson, B. (1999). Why 'encouraging' effect size reporting is not working: The etiology of researcher resistance to changing practices. *The Journal of Psychology*, *133*(2), 133-140.

Vacha-Haase, T. & Nilsson, J. (1998). Statistical significance reporting: Current trends and uses in MECD. *Measurement and Evaluation in Counseling & Development*, *31*(1), 46-57.

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

Table 1
Items which Distinguish Elon University Freshmen from Those at All NSSE Schools (2005)

Item	Elon % Positive	All NSSE Schools % Positive	*d*
Make significantly more class presentations	52	33	0.42
Work more frequently with classmates outside of class	63	43	0.41
Worked w/faculty members on activities other than course work (committees, orientation, student life activities, …)	26	14	0.45
Write significantly more short papers (5 pages or less)	73	39	0.64
Write significantly papers/reports between 5 -19 pages	26	11	0.78
More frequently attend exhibits, galleries, plays, or dances	50	29	0.47
Are more likely to have participated in a learning community than their peers	48	36	0.41
More frequently attend campus events (athletics, special speakers, cultural)	90	56	0.54
Feel support to thrive socially	61	43	0.43
Cognitive/behavioral development: Students contributing to the welfare of their community	67	46	0.45
Are more satisfied about their educational experience	97	87	0.44

Table 2
Items which Distinguish Elon University Seniors from Those at All NSSE Schools (2005)

Item	Elon % Positive	All NSSE Schools % Positive	*d*
Make significantly more class presentations	86	64	0.49
Work more frequently with classmates outside of class	83	59	0.52
Use e-mail more frequently to communicate with their instructors	92	83	0.42
Worked w/faculty members on activities other than course work (committees, orientation, student life activities, …)	42	26	0.50
Write significantly papers/reports between 5 -19 pages	62	38	0.48
Are more likely to have participated in practicum, internship, co-op, field experience, …	91	77	0.58
Are more likely to perform community/volunteer service	93	76	0.52
Are more likely to have studied abroad	73	25	1.35
Are more likely to have a culminating senior experience (capstone, thesis, project, comprehensive exam)	89	65	0.68
Attend campus events (special speakers, cultural performances, athletic events)	82	56	0.62
Are more satisfied about their educational experience	97	88	0.41

Table 3
Items that Do Not Distinguish Elon University Students from Other NSSE Schools (2005)

Item	Class	Elon % Positive	All NSSE Schools % Positive	*d*
To what extent have your exams challenged you?	Freshmen	54%	52%	0.03
	Seniors	50%	53%	-0.04
Course work emphasizes making judgments about the value of information, arguments, or methods	Seniors	78%	72%	0.11