3

Drawing on their experience, current and former NSSE research analysts offer helpful tips and recommendations for institutional researchers on how to analyze student engagement data, including ways to work with multiple years of results and interpret effect sizes.

Analyzing and Interpreting NSSE Data

Pu-Shih Daniel Chen, Robert M. Gonyea, Shimon A. Sarraf, Allison BrckaLorenz, Ali Korkmaz, Amber D. Lambert, Rick Shoup, Julie M. Williams

Colleges and universities in the United States are being challenged to assess student outcomes and the quality of programs and services (McPherson and Shulenburger, 2006; Commission on the Future of Higher Education, 2006). One of the more widely used sources of evidence is student engagement as measured by a cluster of student engagement surveys administered by the Center for Postsecondary Research at Indiana University. They include the National Survey of Student Engagement (NSSE) and its companion projects: the Beginning College Survey of Student Engagement, Faculty Survey of Student Engagement, and Law School Survey of Student Engagement. The University of Texas–Austin hosts the two-year variation of the NSSE, the Community College Survey of Student Engagement.

With more than thirteen hundred colleges and universities using NSSE, many institutional researchers may benefit from guidance about how to understand and use these data. This chapter shares practical tips and recommendations for the analysis and interpretation of NSSE data. We divided the chapter into three parts. The first offers six overarching tips and recommendations for working with student engagement data. We then discuss considerations for the analysis of multiple years of student engagement data. Finally, we describe how effect sizes can be used and interpreted to make student engagement results more meaningful.

NEW DIRECTIONS FOR INSTITUTIONAL RESEARCH, no. 141, Spring 2009 © Wiley Periodicals, Inc.
Published online in Wiley InterScience (www.interscience.wiley.com) • DOI: 10.1002/ir.285

Although we use NSSE data and examples throughout the chapter, institutional researchers can almost always extrapolate the suggestions we provide to other student experience surveys. Instead of technical discussions of such topics as scale construction and factor analysis, we focus on practical, concrete data manipulations and applications for the analytical work of the institutional research professional.

Helpful Tips for Analyzing Student Engagement Data

The primary objective of collecting student engagement data is to discover areas where colleges and universities can improve the quality of the student experience. Student experiences and outcomes vary more among students within an institution than the average student varies between institutions (Kuh, 2003; National Survey of Student Engagement, 2008; Pascarella and Terenzini, 2005). While between-institution analyses emphasize average student performance, within-institution data almost always yield more actionable results. This can be achieved by drilling down into results from subgroups such as men and women or students who participate in certain programs or major in different fields of study.

To help inform improvement efforts, institutional researchers must make engagement results accessible and the reports easy to digest. Toward these ends, this section offers six general recommendations to guide the analysis and reporting of student engagement data to campus leaders and other stakeholders.

Determine the Quality of Institutional Data. When using engagement results to assess the campus experience or inform a particular campus policy, it is necessary to verify that population estimates derived from the data are accurate and precise. In general, data quality is tied to sample size: the more respondents, the more confident one can be in results. Yet short of a 100 percent response rate, no single indicator provides sufficient evidence that a population estimate is truly unbiased. Below we describe the primary data quality measures an institutional researcher should consider in evaluating data quality.

Response Rate. Response rate is the percentage of a sample that completes the questionnaire. Although conventional wisdom holds that the higher the response rate, the better, we encourage a more nuanced exploration of the issue. This conventional view rests on assumptions about nonrespondents. Are they different from respondents and, if so, by how much? *Nonresponse bias,* one potential source of inaccurate population estimates, is a function of both response rate and *nonresponse effect,* the extent to which responders and nonresponders differ on the survey variables of interest (Federal Committee on Statistical Methodology, 2001). Although low response rates may suggest a potential bias in survey estimates of overall population values, they do not necessarily represent bias. As Groves (2006) claims, "There is little empirical support for the notion that low response

rate surveys de facto produce estimates with high nonresponse bias" (p. 670). However, when low response rates are coupled with a nonresponse effect, legitimate concerns about bias are warranted. In fact, even high response rates can result in substantial nonresponse bias when linked with large nonresponse effects. Korkmaz and Gonyea (2008) found only trivial differences between precollege characteristics and academic behaviors among NSSE responders and nonresponders, suggesting that nonresponse bias may be trivial. Studies of nonresponse bias by NSSE in 2001 and 2005 also concluded that nonresponse effects are minimal (National Survey of Student Engagement, 2008). Still, as with all other survey results, institutions may vary greatly in terms of nonresponse effects and bias. So under certain conditions, very low response rates may render the results problematic, especially after careful consideration of other data quality measures.

Sampling Error. According to Salant and Dillman (1994) sampling error is a fact of life for those using survey data. Sampling error occurs when respondents represent a subset, or sample, of the total population. It estimates how much respondents could differ on survey measures from the entire population of students at an institution. For example, if 60 percent reply "very often" to a particular item and the sampling error is estimated at plus or minus 5 percent, then the actual population value is likely to be between 55 and 65 percent. Estimating sampling error is a function of the number of students who responded to the survey (n) and the total number of students in your population (N) (see equation 3.1):

$$e = \sqrt{\frac{.9604(N-n)}{n(N-1)}} \tag{3.1}$$

Increasing the number of respondents relative to the total population reduces sampling error. Smaller sampling errors such as ±3 or 5 percent are preferred, although data with larger sampling errors (such as ±10 percent) need not be dismissed offhand but rather interpreted more conservatively.

Sampling error is based on the population of interest. Therefore, to estimate the sampling error for first-year male students, one must base the analysis on the number of all first-year male students in the population and the number of first-year male respondents.

Proportional Representation. It is also necessary to determine the extent to which respondent demographics match those of the population. If students with certain characteristics make up 70 percent of the campus population but only 40 percent of the survey respondents, researchers may need to make adjustments, especially if that variable is related to engagement. Weighting or other statistical procedures may counter the potential biases in the data. For instance, NSSE weights its data by gender and enrollment status, not only because women and full-time students respond at higher rates, but also because they respond differently to important NSSE measures. Weighting also helps determine whether changes from year to year

can be linked to such things as changing student demographics or campus initiatives.

Collapse Response Categories for Reporting and Analysis. Institutional reports of student experiences typically are too detailed to pass along to campus leaders. A task of the institutional researcher is to extract from the raw data and basic reports the most meaningful and relevant pieces of information. One particularly useful approach is to collapse the response categories for individual items into fewer categories in order to succinctly convey results (Table 3.1). For example, the response set "never, sometimes, often, very often" can be recoded so that "very often" and "often" are combined into a new category labeled "frequently." Other times you may simply want to examine the percentage of students who report that they "never" do something.

This approach works when using descriptive analyses to identify percentage differences between subgroups. This technique also is instructive when doing more sophisticated statistical analyses, such as logistic regression to predict students who will participate in high-impact activities (Kuh, 2008).

By collapsing response options using the suggestions in Table 3.1 or other similar methods, institutional researchers can more easily review the results and look for interesting findings to present to decision makers.

Combine Questions into Workable Scales. As with any other set of behavioral or attitudinal constructs, student engagement cannot be measured with a ruler or thermometer as we would with some physical characteristics. As a result, questionnaires often include series of questions to gauge student behaviors and attitudes. Yet the responses to a single question may be too narrow for decision making on broader policies, while results from dozens of individual questions may be mixed and inconclusive. Thus, it is useful to combine individual items into scales that consist of a limited number of conceptually related questions. Scales reduce the number of variables in analytical models, may have better reliability, and ultimately may convey more meaningful information than individual questions.

In addition to the five benchmarks of effective educational practice (see Chapter One, this volume, for a description of the benchmarks), NSSE analysts confirmed a deep learning scale of twelve questions with three subscales (Nelson Laird, Shoup, and Kuh, 2006; Nelson Laird, Garver, Niskodé-Dossett, and Banks, 2008). The NSSE instrument also contains three self-reported gains scales based on sixteen questions about the extent that the student's experience at the institution contributed to his or her learning and growth. In addition, Pike (2006a, 2006b) identified eleven useful "scalelets." (SPSS syntax for creating these NSSE scales can be found at www.nsse.iub.edu/html/syntax_library.cfm.)

Computing Scale Scores. The simplest method to compute the scale score is to sum the response values for each of the individual items. For example,

Table 3.1. Collapsing Response Options with NSSE Data

New Category	*Original Response(s) Used*	*Sample Question*	*Recommended Uses*
Frequently	"Very often" and "often"	How often have you asked questions in class or contributed to class discussions?	Pinpoint activities students do most on campus; look at percentage differences among subgroups.
Never	"Never"	How often have you made a class presentation?	Identify possible areas in need of improvement.
Substantial	"Very much" and "quite a bit"	How much has your course work emphasized memorizing facts or ideas?	Examine the amount that course work emphasizes higher-order learning activities, gauge aspects of the campus environment, and look at self-reported gains.
Done	"Done"	Have you done or do you plan to study abroad before you graduate?	Report how many students participate in high-impact practices; break down by academic major.
Friendly or helpful	Combine top three positive responses on a scale of 1 to 7	What is the quality of your relationships with other students [or with faculty members]?	Summarize students' relationships with key campus groups—students, faculty, and administrators.
Sixteen or more hours	Combine all responses greater than "11–15" hours	How many hours do you spend in a typical seven-day week preparing for class?	Collapse responses for each item differently to create the most useful responses.
Quartiles, or above- and below-average groups	All benchmark values recoded into four equal groups, or two groups divided by the mean	Benchmarks, scales, or other continuous measures	Investigate if certain subgroups are over- or underrepresented. For example, if 20 percent of the top quartile is male but males are 40 percent of the entire sample, you can claim male underrepresentation in this high-performing quartile.

when combining five items that are coded 1 = never, 2 = sometimes, 3 = often, and 4 = very often, the sum will have a minimum score of 5 and a maximum score of 20. A better option may be to compute the mean score for these items, so that the scale score remains within the original range (1 to 4) and can be interpreted according to original response units. For example, a mean score of 3.1 is about an average response of "often."

In some situations, a researcher may need to combine items from different response sets and value ranges. For example, two items might have a range of 1 to 4, and three items might have a range of 1 to 7. If one simply sums or averages the response values to score the scale, the questions with the greater number of options will have a larger influence on the overall score. To balance the contribution of the individual questions, standardize them with a mean of 0 and a standard deviation of 1, and then sum the standard scores. Another option is to recode the individual response values into a common scale range like the NSSE research team does in creating the benchmark, deep learning, gains, and other scale scores. To do so, NSSE converts each item into a scale of 0 to 100, an arbitrarily chosen range for reporting purposes, using equation 3.2:

$$[(\text{response value} - 1)/(\text{total number of response values} - 1)] * 100 \quad (3.2)$$

For example, an item with four response options is converted to 0, 33.3, 66.7, and 100. Afterward, compute the scale score by computing the mean of the mean of these recoded items.

When two scales share a common item (or set of items), be sure not to enter them into the same statistical models or equations. The common item or items will produce artificially higher correlations and confound the statistical analysis.

Compute Basic Statistical Comparisons Against Published Norms. In this section, we discuss how institutional researchers can analyze their data for statistical differences against published aggregates and what these findings mean in a practical sense. Although more advanced methods are available, the approaches presented here are offered as shorthand calculations to produce results with a few simple computations.

Statistical Difference: Calculating t*-Tests.* The *t*-test determines whether the means of two groups are statistically different from one another, that is, the likelihood that the difference between groups occurred by chance alone. To calculate a *t*-test, one just needs a few descriptive statistics. Equation 3.3 can be used to calculate a *t* score, where M_1 is the mean score for the selected institution, M_2 is the mean score for the comparison group, and *SEM* is the standard error of the mean:

$$t = \frac{M_1 - M_2}{SEM} \quad (3.3)$$

This definition of SEM is used in one-sample *t*-tests that consider the comparison group to be a population parameter, not a sample estimate. With the exception of extremely small sample sizes, *t* values greater than 2.0 can be interpreted to mean that a statistically significant difference exists at the $p < .05$ level. Similarly, *t* values greater than 2.6 imply significance at the $p < .01$ level, and *t* values greater than 3.3 imply significance at the $p < .001$ level:

For example, an institution may be concerned about the quality of relationships between students and faculty, a NSSE item that is coded on a seven-point rating scale from 1 = "unfriendly, unsupportive, sense of alienation" to 7 = "friendly, supportive, sense of belonging." At the institution, 466 seniors answered this question with a mean of 5.7 and SEM of .06. The mean for all seniors who completed this item in 2008 was 5.4, which can be obtained from the NSSE's Web site. The statistical significance can be determined by calculating the *t* value as follows.

$$t = \frac{M_1 - M_2}{SEM} = \frac{5.7 - 5.4}{.06} = 5.31$$

Because the *t* value is greater than 3.3, it indicates that the institutional mean is significantly greater than the NSSE average at the $p < .001$ level.

Practical Difference: Calculating Effect Size. An effect size, considered a measure of practical significance, is any measure of the strength of the relationship between two variables. For the purposes of this chapter, we use Cohen's *d* (Cohen, 1988), the difference between two means divided by the standard deviation of a comparison (or norm) group (or, alternatively, the pooled standard deviation of the two groups). The *d* statistic expresses the mean difference in standard units and can be interpreted in relative terms. Later in this chapter, we discuss the interpretation of effect sizes in more detail.

Computing an effect size can be particularly helpful for institutions and comparison groups with large sample sizes. Significance tests can be problematic with studies that include large sample sizes because as sample size increases, standard errors of the mean decrease, and thus significance tests more easily yield higher *t* values. The effect size does not have this limitation, so all comparisons, even those with large sample sizes, can be interpreted in the same general way from a practical standpoint. To calculate effect sizes between an institution's results and a published norm, use equation 3.4, where M_1 is the institutional mean, M_2 is the norm group mean, and *SD* is the standard deviation of the norm group:

$$d = \frac{M_1 - M_2}{SD} \tag{3.4}$$

Using the same example as above, the institution's mean for the quality of student relationships variable for seniors was 5.7, the NSSE average for all

senior participants in 2008 was 5.4, and the standard deviation for all senior participants in 2008 was 1.4. Applying equation 3.3, the effect size is:

$$d = \frac{M_1 - M_2}{SD} = \frac{5.7 - 5.4}{1.4} = .22$$

Thus, while the *t*-test revealed that the quality of student relationships with faculty members at the institution was significantly higher than the NSSE average, the effect size shows that the difference is substantive but perhaps small.

Test for Interaction Effects. When looking at student differences within survey results, the number of subgroups can be daunting, and drilling down can create too many results to be practically considered in policy development. In addition, dividing the data by a large number of subgroups has the potential to diminish the statistical power of the analysis. To avoid some of these pitfalls, researchers can test for the existence of interaction effects, variously termed *conditional, joint, contingency,* or *moderating effects.* Pascarella and Terenzini (2005) emphasized the importance of studying conditional effects and urged scholars to take such effects into account in future research. As illustrated in Figure 3.1, an interaction is present when the association between two variables (X_1 and Y) depends on changes in a third variable (X_2) (Agresti and Finlay, 1999).

For example, to see if an interaction exists in the effects of gender and majoring in a science, technology, engineering, and mathematics (STEM) field on student-faculty interaction, one may chart the means of each subgroup (male/female, and STEM/non-STEM). It appears in Figure 3.2 that the effect of gender on student-faculty interaction is different for STEM and non-STEM majors at this institution. Non-STEM females are more likely than their male counterparts to interact with faculty, while the inverse is true in the STEM fields.

Test for Significant Interactions. When the two independent variables are categorical as shown in Figure 3.2, the analysis of variance (ANOVA) procedure includes a test of interaction effects among the independent variables. When one or both of the independent variables are continuous, enter-

Figure 3.1. Interaction Effect Model

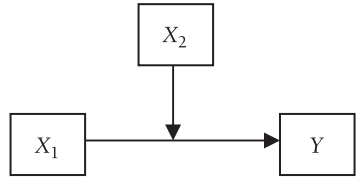
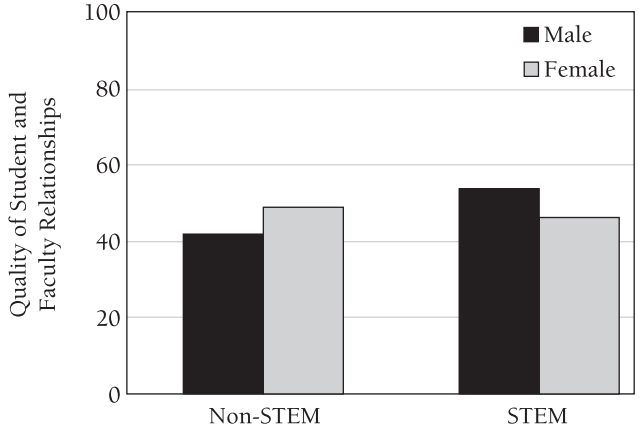


Figure 3.2. Student-Faculty Interaction by Gender and Discipline at One Institution



ing cross-product terms into a regression model can test for interactions. For example, if the hypothesis is that the effect of frequent conversations with diverse individuals (X_1) on the quality of student relationships (Y) depends on the student's race (X_2), one can test the hypothesis by creating a cross-product term (for example, $X_1 * X_2$) that represents the possible interactions between race (dummy coded) and conversations with diverse others. This is done by multiplying each of the dummy race variables by the "conversations" variable.

To illustrate, we will look at just two race categories, African American and white, where the African American variable (X_2) is entered into the model and white is left out as the reference group. The interaction term ($X_1 * X_2$) is entered separately at the end of the regression model so the researcher can look for a significant change in the amount of variance explained (change in R^2). If the change is significant, the interaction term explains at least some additional variance, and the main-effects-only model may not be sufficient. In the model without the interaction term, the coefficients represent the effects on the dependent variable while taking all other independent variables into account. With an interaction term in the model, the interpretation changes so that the coefficient is the effect on the dependent variable when the other independent variable is equal to one (Jaccard and Turrisi, 2003). In other words, the coefficients for the variables represent conditional relationships, not main effects. These conditional effects can be calculated by substituting zeros for the dummy variables and creating separate regression equations for each subgroup. This process has the same effect as fitting separate regression lines for each group.

Table 3.2. Regression Coefficients for Conversations with Diverse Others, Being African American, and Interaction Term on Quality of Student Relationships

	Parameter	*Estimate*
α	Intercept	–.02
x_1	Conversations with diverse others	.20
x_2	African American	–.06
$(x_1 * x_2)$	Conversations * African American	.09

Using data from one NSSE institution, Table 3.2 presents illustrative model coefficients, including the interaction term effect. Because these coefficients are conditional and not main effects, separate equations for each race are computed (Table 3.3) (white is indicated when the African American variable is set to zero). The nonparallel slopes in Figure 3.3 indicate that more frequent conversations with diverse others have a stronger effect on the quality of student relationships for African American students than it does for white students.

Examining interaction effects provides administrators and faculty members more specific information as to those students who are more or less engaged and the degree to which they develop desired skills and competencies.

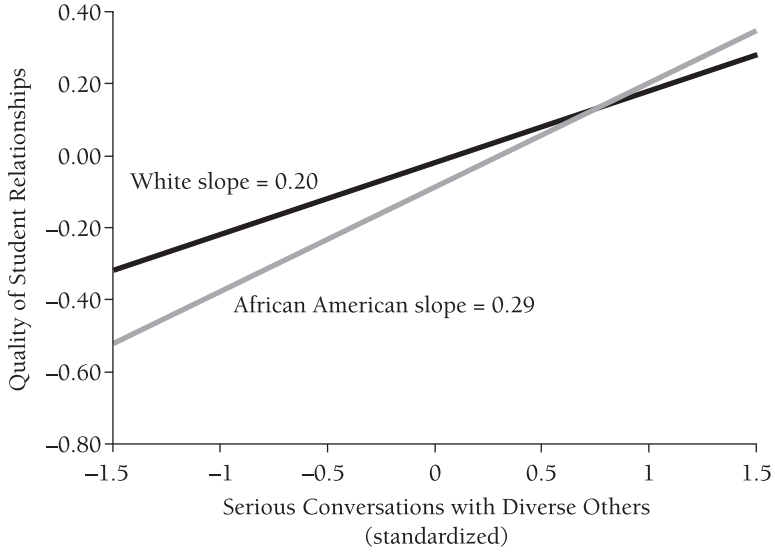
Link Engagement Data to Other School Records. All surveys are limited in the amount of information they provide, so it can be fruitful to link survey data with other student records for a more comprehensive understanding of student engagement and learning on campus. To make this possible, NSSE returns data to institutions with unique student identifiers that institutions provided beforehand in their student population files.

For example, institutional researchers can link high school grades and standardized test scores typically collected during the admission process with students' NSSE responses to determine whether first-year students from different educational backgrounds engage in educationally purposeful activities in comparable ways. While students who achieved lower grades

Table 3.3. Summary of Prediction Equation Allowing Interaction for the Effects of Serious Conversations on Quality of Student Relationships

Race	*Y Intercept*	*Slope*	*Regression Equation*
White	–.02	.20	$Y = –.02 + .20x$
African American	–.02 – .06	.20 + .09	$Y = –.08 + .29x$

Figure 3.3. Race Variations in the Effect of Serious Conversations on Supportive Campus Environment



in high school may be less well prepared for college-level work, it may concern faculty and staff to learn that such students also spend fewer hours studying, read less, and are less likely to write multiple drafts of a paper, all of which also could lead to poor performance in college. If such a gap is known, institutions can take action to promote a better understanding of the amount and type of work that is expected for success in the first year. They may offer additional academic enrichment sessions during orientation for students more likely to struggle academically, emphasize good practice in writing and reading in entry-level first-year classes, and reach out to students with lower high school achievement through academic support service centers.

Institutional researchers can also link data about persistence, financial aid, and student satisfaction with survey results to evaluate the effectiveness of specific academic departments or divisions, special academic programs, and student affairs activities. For example, one could examine the first-year engagement patterns of NSSE respondents who reenroll as sophomores to determine which characteristics or activities were associated with persistence. If working more than twenty-five hours per week off campus has a negative relationship with persistence, administrators might reexamine the number of work-study positions available on campus, review how scholarship dollars are distributed, and caution students through orientation and advising activities against working too many hours while enrolled.

Table 3.4. Examples of Student Records to Link with Engagement Data

Source	*Description*	*Sample Questions for Analysis*
Admissions	High school grades, standardized test scores, high school attended	Do students of varying academic backgrounds report similar levels of academic effort?
Financial aid	Scholarship award designations, student need	Are elite academic scholarship winners engaged in more deep learning and research with faculty?
Orientation, first-year experience	Learning community or first-year seminar assignments	How does first-year seminar participation relate to ratings of the campus environment?
Registrar	Student enrollment, grades, progress toward degree, transcripts	Which engagement activities are correlated with continued enrollment and with academic success?
Academic support center	Student use of support services	Do students who visit the writing center report greater deep learning and general education gains?
Testing center	Placement test results	How do provisional students rate the campus environment for learning?
Athletics	Team participation	Are student athletes less engaged than nonathletes in effective educational practices?
Academic departments	Honors programs, capstone courses, portfolios	Do students in selective academic programs participate in more effective educational practices?
Student affairs	Cocurricular participation (for example, fraternity or sorority, student government)	Are students who participate in cocurricular activities as engaged in academic learning?

Table 3.4 lists examples of student records that may be linked with engagement data and suggests questions researchers may ask. Such analyses can also generate discussions that contribute to the development of a culture of evidence. Using multiple sources of data expands the number of questions that institutional researchers can answer for campus assessment, particularly to inform critical programmatic and curricular decisions.

Working with Multiple Years of Student Engagement Data

To inform institutional improvement efforts, it is best to collect data about student experiences across multiple years. More than three-quarters of NSSE

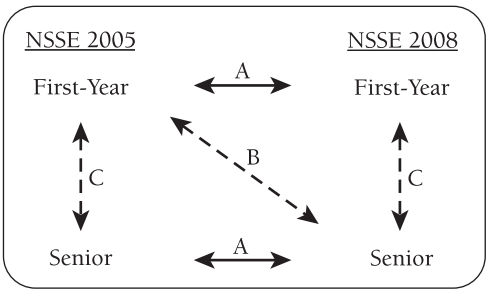
institutions have participated more than once. The UCLA-based Cooperative Institutional Research Programs has hundreds of institutions that have participated annually for decades. As colleges and universities collect multiyear data on student experiences, institutional researchers seek ways to document and record changes, track possible trends, and evaluate specific campus initiatives. In this section, we recommend analytical approaches to multiyear student engagement data, again using NSSE as the primary example.

Multiyear student engagement data can answer many important questions. The arrows in Figure 3.4 represent three scenarios for an institution that collected NSSE data from their first-year and senior students in 2005 and 2008. We call them (A) cohort comparisons, (B) longitudinal analyses, and (C) cross-sectional analyses.

Cohort Comparisons. This preferred approach, represented by the arrows labeled A in Figure 3.4, compares the engagement of students in a baseline year with the engagement of students at the same class level in a later year or years. With cohort comparisons, it is assumed that each year's data offer the best estimate of the class-level population of students during that year. Of course, the first-year and senior cohorts include a different sample of students for each year in the data, and the students may differ in their demographic and background characteristics. Analysts can make statistical adjustments for such differences and can help to determine if engagement in educationally effective learning practices during the first-year or the senior-year experience has changed over time.

Longitudinal Analysis. Represented by the arrow labeled B in Figure 3.4, longitudinal analysis tracks a panel of students from the first year to the senior year. The advantage of this approach is that unlike the cohort comparisons, one can observe the same students over time, which means that such characteristics as gender, ethnicity, family background, and precollege experiences are constant. Thus, interpretation of results allows the researchers to focus on environmental factors that may influence the nature and frequency of student engagement in various areas.

Figure 3.4. Approaches to Multiyear Analysis



One serious drawback of longitudinal analysis is the inevitable attrition of study participants, which may compromise data quality and limit the conclusions that can be drawn from the data. For various reasons, only a portion of first-year respondents will persist to the senior year, and many of them take different amounts of time to do so. Even those who persist to the senior year may not respond to the survey each year they are invited. In addition, panel data containing only students who entered as first-year students and persisted to the senior year exclude the nonpersisters and the large cohort of transfer students who enrolled after the first year.

Cross-Sectional Analysis. Cross-sectional analysis, represented by the arrows labeled C in Figure 3.4, compares the engagement of first-year students and seniors from the same year. For reasons similar to those mentioned above, we generally discourage this approach. Cross-sectional analysis includes both nonpersisters in the first year and transfers in the senior year; thus, the types of students defining each cohort are different. Unlike the longitudinal approach, the cross-sectional analysis attempts to draw conclusions about college impact from two cohorts containing dissimilar students at disparate stages of their college careers.

Another limitation of cross-sectional or longitudinal approaches is introduced when attempting to estimate college impact or value-added by comparing first-year and senior results. This intended purpose is problematic because engagement data represent process indicators, not tests of content acquisition or achievement. As such, a student's engagement score is not necessarily expected to grow or increase from one year to the next, but rather is an estimate of the student's experiences within the context of the courses and campus environment at the time. A host of academic and social variables can affect the quality of student experiences between the first and senior years of college. For example, the first year of college may differ from the senior year in the type of courses students take, class sizes, and the nature of relationships with faculty members. Student living arrangements differ, as do their peer networks, participation in clubs and other activities, and level of intellectual and personal development. Consequently seniors are more likely to interact with faculty members more often about career plans, while first-year students may have more frequent conversations with peers from diverse backgrounds. Ceiling effects may also have a bearing. For example, if the students scored fairly high in a particular area of engagement in both the first year and the senior year, this may be a very positive finding even though negligible differences exist between performance at the two points in time.

Given these considerations, four steps are necessary to analyze multiyear data effectively: (1) identify and focus on specific questions, (2) employ appropriate analytical methods, (3) verify data quality for each year in the analysis, and (4) create a multiyear data file.

First, institutional researchers should identify and focus on specific questions. Multiyear research questions should be specific, answerable, and relevant to pressing campus issues. Tying them to strategic priorities, initia-

tives to be evaluated, or policy decisions increases the study's chances of being well received and actually used. The following three research questions represent typical approaches to multiyear analysis:

1. Is the level of academic challenge reported by our students about the same in 2008 as it was in 2006?
2. Given the implementation of initiative X in 2006–2007, how much did our students' level of active and collaborative learning increase from 2006 to 2008?
3. Given several campus initiatives aimed at increasing contact between faculty and students over the past decade, what trends are apparent in the frequency our students have interacted with faculty?

The first question is about stability, which can also be an indicator of data reliability. In the absence of a major campus initiative or shift in the characteristics of the student population, one would expect an institution's student engagement results to be relatively stable from one year to the next. The second question seeks to gauge whether a new program or initiative is associated with higher levels of student engagement. With data collected prior to and after the implementation of an initiative or program on campus, a multiyear analysis can mimic a pretest-posttest research design by comparing engagement levels before and after implementation. Finally, the third question focuses on trends. The number of data points needed for a trend analysis is subjective and in part relies on institutional context and the questions being investigated.

Second, institutional researchers should employ appropriate analytical methods. Once the research questions have been determined, the researcher needs to choose analytical methods that will help identify statistically significant and meaningful changes from year to year. Space allows us only to suggest some ideas in this regard, and not an in-depth discussion or a comprehensive guide for using these procedures. Keep in mind that these methods are not mutually exclusive and can be used together to test year-to-year changes in results.

T-tests are one initial step to determine if statistically significant differences exist and to identify which student experiences may have changed between the years. Although best used with interval-level data, as with NSSE benchmark scores, many use this robust test (and other similar tests) with ordinal-level data.

ANOVA is a good method for reviewing a large number of variables for which three or more years of data exist. A finding of no significant difference between years suggests that student experiences may not have changed across the years. Alternatively, if this test identifies significant differences between years, post hoc tests help identify where changes have occurred.

Regression can also address questions regarding significant changes between years. By dummy-coding the year variable in the model and using

the base year as the reference group, one can test for significant differences between the base year and subsequent results. If no statistically significant coefficients emerge, one may conclude that results are unchanged relative to the base year. In addition, assess trends by analyzing the size of the "year" coefficients in sequential order; coefficients may consistently increase from year to year, decrease, or stay stable. Furthermore, using regression analysis with student characteristics as controls provides more detailed information for identifying variables associated with year-to-year changes. Using statistical controls makes it possible to attribute score differences to a campus initiative or policy change rather than shifting student demographics.

Because large sample sizes are often tied to statistical significance even when differences are trivial, effect size estimates reflect practical significance because they indicate the relative magnitude of the difference. For example, Cohen's *d,* the effect size provided on NSSE reports, expresses the mean difference in standard units that can be interpreted regardless of large samples. With regression models, one can standardize dependent variables so that the coefficient of the dummy-coded year variable (as described in the section on regression analysis) can be read as an effect size. Using effect size statistics requires the researcher to establish criteria for determining whether a meaningful change has occurred. The next major section of the chapter has more discussion on this topic.

Analyzing collapsed response percentages by year (see Table 3.1) may make it possible to establish criteria to evaluate whether the percentage change is meaningful. For example, a minimum increase of 5 percent within a two-year period may indicate a real positive change. Because this approach does not test for statistical significance and uses a more subjective evaluation, the researcher may want to take into account sampling error statistics to establish the criteria. The greater the sampling error for each year, the more conservative one should be with establishing criteria for change.

Third, institutional researchers should verify the data quality for each year in the analysis. Reviewing the quality of the data beforehand is especially critical with multiyear studies because each year of data employed contains a certain amount of error. Some survey administrations yield more precise population estimates than others. In some years, the institution may have better data quality due to a higher response rate or because they intentionally oversampled. In other years, there may be student groups that are overrepresented more than others.

Fourth, institutional researchers then create the multiyear data file. Preparing a multiyear data set includes identifying variables that have not changed over the years and merging the cases from all years into a single file. Even minor changes in item wording or the order that items appear on the questionnaire can affect how individuals respond (Sudman, Bradburn, and Schwarz, 1996). NSSE makes available an Excel spreadsheet that tracks every variable by year, detailing whether the item has changed and if it can be compared over time (see www.nsse.iub.edu/html/researchers.cfm and

select "NSSE Survey Instruments"). This is especially important if the institution relies solely on the reports generated by NSSE or another provider. If an item has been altered from one year to the next, results for that item on the new report may not be comparable with the same item on the older report. Merging multiyear data can be a tedious job, but doing this carefully will ensure accurate results.

Putting Results into Context: Interpreting the Effect Size

Earlier we recommended a straightforward way to compare an institution's students with those of a published norm group by computing the Cohen's *d* effect size. In this section, we address a frequently asked question about how to interpret effect sizes in the context of actual student engagement results. How big an effect is .3 or .6? Intentionally vague about precise cut points and decision rules, Cohen (1988) reluctantly defined effect size as "small, $d = .2$," "medium, $d = .5$," and "large, $d = .8$" and urged researchers to interpret effect size based on the context of the data. Nevertheless, researchers have widely accepted and incorporated Cohen's definition of small, medium, and large into many social science studies.

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. In terms of NSSE, the vast majority of effect sizes on its benchmark comparison reports were either trivial (less than .20 in magnitude) or small (.20 to .49 in magnitude) by Cohen's definition.

Following this logic, the NSSE research team considered ways in which benchmark differences would be visible in the data. We compared students attending different sets of institutions according to their performance on the NSSE benchmarks and constructed model comparisons to resemble effect sizes of increasing magnitude. For example, we posited that a small effect size would look like the difference between institutions in the second quartile compared with institutions in the third quartile of the distribution of all institution-level NSSE benchmark scores. Likewise, a medium effect would look like the difference between institutions in the lower half and institutions in the upper half of the distribution. A large effect would be like comparing institutions in the lowest quartile and those in the highest quartile. Finally, a very large effect would resemble the difference between institutions in the top 10 percent and bottom 10 percent.

As a result, the effect sizes for these small, medium, large, and very large model comparisons turned out to be fairly consistent from one benchmark to the next, so we recommended a somewhat finer grained approach to effect size interpretation than Cohen's definition, shown in Table 3.5. Because we based these calculations on benchmark distributions, the new reference values are proposed for NSSE benchmark comparisons and not for individual item mean comparisons. Like Cohen's, one should not interpret

Table 3.5. Proposed Reference Values for the Interpretation of Effect Sizes from NSSE Benchmark Comparisons

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

these values as precise cut points, but rather as a coarse set of thresholds or minimum values by which to consider the magnitude of an effect size.

As expected the majority of effect sizes based on these new reference points were trivial (less than .1), small, and medium—a finer distribution within categories from what we saw based on Cohen's definitions. Approximately one-quarter to one-third of all effect sizes appear to be in the trivial range, more than 40 percent are considered small, and the new medium range captures about 20 to 25 percent of all effect sizes. Large and very large effect sizes are relatively rare.

Finally, with effect size analysis, we recommended another important step for putting benchmark comparisons in context: examine individual item responses to see what student behaviors or institutional conditions may be associated with the result. In this instance, looking at the frequency reports can help make benchmark scores and effect sizes more accessible and understandable. For example, many combinations of individual item results can produce a particular effect size. Consider two institutions with the same effect size on a particular benchmark. The first may have large percentage differences on just a few of the benchmark items, while the second could have small percentage differences on all of the items. A series of small differences can accumulate into appreciable effect sizes when combined to form the benchmark score. So looking at the response frequencies of the items within the scale can provide an instructive explanation for a statistical comparison or effect size, which can help administrators and policymakers focus on specific action plans to improve the undergraduate experience. (A more comprehensive discussion of this approach can be found at www.nsse.iub.edu/pdf/effect_size_guide.pdf.)

Conclusion

Today's higher education environment demands of accountability, transparency, and public reporting are more than enough to keep institutional researchers busy. The widespread use of student experience surveys includ-

ing NSSE adds another important layer of information about student and institutional performance. While some institutions have not, or may never, use NSSE, almost all likely have information about student experiences, whether from a different nationally published instrument or from a locally developed tool. We hope this chapter provides some constructive ideas to help institutional researchers make the best of student engagement results that campus leaders and others will use to improve the quality of the undergraduate experience.

References

Agresti, A., and Finlay, B. *Statistical Methods for the Social Sciences.* (3rd ed.) Upper Saddle River, N.J.: Prentice Hall, 1999.

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences.* (2nd ed.) Mahwah, N.J.: Erlbaum, 1988.

Commission on the Future of Higher Education. "A Test of Leadership: Charting the Future of U.S. Higher Education." Washington, D.C.: U.S. Department of Education, 2006. Retrieved Sept. 4, 2008, from http://www.ed.gov/about/bdscomm/list/hiedfuture/reports/final-report.pdf.

Federal Committee on Statistical Methodology. *Measuring and Reporting Sources of Error in Surveys.* Washington, D.C.: U.S. Office of Management and Budget, 2001.

Groves, R. M. "Nonresponse Rates and Nonresponse Bias in Household Surveys." *Public Opinion Quarterly,* 2006, *70*(5), 646–675.

Jaccard, J., and Turrisi, R. *Interaction Effects in Multiple Regression.* (2nd ed.) Thousand Oaks, Calif.: Sage, 2003.

Korkmaz, A., and Gonyea, R. M. "The Effect of Precollege Engagement on the Likelihood of Response to the National Survey of Student Engagement." Paper presented at the Annual Forum of the Association for Institutional Research, Seattle, May 2008.

Kuh, G. D. "What We're Learning About Student Engagement from NSSE." *Change,* 2003, *35*(2), 24–32.

Kuh, G. D. *High-Impact Educational Practices: What They Are, Who Has Access to Them, and Why They Matter.* Washington, D.C.: Association of American Colleges and Universities, 2008.

McPherson, P., and Shulenburger, D. *Toward a Voluntary System of Accountability Program (VSA) for Public Universities and Colleges.* Washington, D.C.: National Association of State Universities and Land-Grant Colleges, 2006. Retrieved Sept. 27, 2008, from http://www.voluntarysystem.org/docs/background/DiscussionPaper3_Aug06.pdf.

National Survey of Student Engagement. "NSSE 2008 Psychometric Properties." 2008. Retrieved Dec. 1, 2008, from http://www.nsse.iub.edu/2008_Institutional_Report.

National Survey of Student Engagement. *Promoting Engagement for All Students: The Imperative to Look Within.* Bloomington: Center for Postsecondary Research, Indiana University School of Education, 2008.

Nelson Laird, T. F., Garver, A. K., Niskodé-Dossett, A., and Banks, J. V. "The Predictive Validity of a Measure of Deep Approaches to Learning." Paper presented at the Annual Meeting of the Association for the Study of Higher Education, Jacksonville, Fla., Nov. 2008.

Nelson Laird, T. F., Shoup, R., and Kuh, G. D. "Measuring Deep Approaches to Learning Using the National Survey of Student Engagement." Paper presented at the Annual Forum of the Association for Institutional Research, Chicago, May 2006.

Pascarella, E. T., and Terenzini, P. T. *How College Affects Students: A Third Decade of Research.* San Francisco: Jossey-Bass, 2005.

Pike, G. R. "The Dependability of NSSE Scalelets for College- and Department-Level Assessment." *Research in Higher Education,* 2006a, *47*(2), 177–195.
Pike, G. R. "The Convergent and Discriminant Validity of NSSE Scalelet Scores." *Journal of College Student Development,* 2006b, 47(5), 551–564.
Salant, P., and Dillman, D. A. *How to Conduct Your Own Survey*. New York: Wiley, 1994.
Sudman, S., Bradburn, N. M., and Schwarz, N. *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass, 1996.

PU-SHIH DANIEL CHEN is assistant professor of higher education at the University of North Texas.

ROBERT M. GONYEA is associate director of the Center for Postsecondary Research at Indiana University–Bloomington.

SHIMON A. SARRAF is a research analyst with the Center for Postsecondary Research at Indiana University–Bloomington.

ALLISON BRCKALORENZ is a research analyst with the Center for Postsecondary Research at Indiana University–Bloomington.

ALI KORKMAZ is a research analyst with the Center for Postsecondary Research at Indiana University–Bloomington.

AMBER D. LAMBERT is a research analyst with the Center for Postsecondary Research at Indiana University–Bloomington.

RICK SHOUP is a research analyst with the Center for Postsecondary Research at Indiana University–Bloomington.

JULIE M. WILLIAMS is a research analyst with the Center for Postsecondary Research at Indiana University–Bloomington.