A Potential Barrier to Completing the Assessment Feedback Loop

Ed Furlong, Promod Vohra
Northern Illinois University

Abstract

Northern Illinois University’s College of Engineering and Engineering Technology employs a comprehensive nine-component assessment model. Each element in the assessment model (Pre-test, Post-test, and Portfolio; Standardized Testing; Student and Faculty Surveys; Student Internships and Cooperative Work Performance; the Capstone Experience; Student Placement Information; Employer Surveys; Alumni Participation; and Peer Review of the Curriculum) provides a mechanism for data collection.

Within the context of our assessment model, this paper details strategies for analyzing and using assessment results as feedback directed toward the improvement of total program quality. Incorporating feedback into the assessment process is often difficult. Assuming the measurement of selected learning outcome criteria is both valid and reliable, benchmarks for acceptable performance must be established and decision rules that provide a basis for detecting meaningful differences must be formulated. And these tasks are conducted in a policy environment where the implementation of affirmative steps may be constrained by numerous internal and external stakeholders.

One of the most fundamental problems with assessment research involves how assessment results are to be placed within a meaningful comparative context. Any analysis of assessment results involves ascertaining the significance of differences from an established performance baseline, a performance goal, or other criteria. The significance of any comparisons that are made may be evaluated using statistical and/or substantive criteria. This paper will explore the potential and limits of statistical analysis, particularly as both relate to the concept of statistical power in survey research, and discuss several strategies for dealing with the problems posed by inadequate numbers of respondents.

Introduction

In every academic program, teaching practices and student learning have always been important issues to consider. However, the overall thrust of assessment has changed markedly. During the past decade, the focus of assessment has shifted from input variables (resources committed), to output variables (counting students and graduates), to learning and performance outcomes (highly specific criterion-related competencies). And as is often the case, academic programs, notorious for their conservatism and procrastination, have been compelled to shift their assessment efforts from inputs to outcomes due to the higher levels of accountability demanded by external stakeholders such as university governing bodies, accreditation agencies, state boards of higher education, and the industrial consumers of trained students. The groundswell for increased accountability in higher education has also filtered upward from concerned students and parents, to interested state legislators, state legislatures, and executive branch agencies from...
all levels of the federal system. Significantly, the practice-based emphasis of engineering and engineering technology programs has made accountability issues even more relevant for them than other academic disciplines. Thus, there is a strong need for educational resources and priorities to be directed toward establishing and maintaining compound assessment models that have the institutional capacity to measure learning outcomes in ways which fulfill the expectations of external stakeholders who frequently control the accreditation status or budgets of the academic program.

The ABET evaluation of engineering programs represents only a fraction of the ongoing evaluation effort connected with the overall operation of a college. State boards of higher education, regional accreditation agencies, and layers of university-wide program review committees are also assessing learning outcomes, policies related to specific issues, and overall college operations. In order to justify their efforts to their funding agencies and to demonstrate the integration of assessment into their teaching models, academic programs have quickly developed assessment models that are theoretically sound and effective. Even so, the ABET accreditation is the most renowned and distinguished credential that any undergraduate engineering program may aspire to. And, today, in the increasingly competitive marketplace of higher education, ABET accreditation is a must for any engineering program.

Like other colleges, Northern Illinois University’s (NIU) College of Engineering and Engineering Technology (CEET) has also developed a comprehensive nine component assessment model that incorporates assessment activities ranging from course assessment to standardized testing to obtaining feedback from participants such as students, faculty, alumni, employers, industry etc. The college has been actively gathering data from various sources over the past three years. Once this assessment data is collected, the task shifts toward accurately analyzing a large mass of empirical evidence from myriad sources. During this process, trends are noted and inferences with policy implications are made. These findings are then incorporated back into the system as feedback so that the engineering programs in the college may continuously strive to improve their programs in meaningful ways. But roadblocks exist, and the purpose of this paper is to disseminate some useful strategies we learned in the “college of hard knocks” to our colleagues.

The Neglected Importance of Power Analysis in Assessment

When doing survey research to assess student learning outcomes, a standard analysis involves the comparison of the mean responses of similarly situated but different participants along affective or behavioral dimensions. We may, for example, wish to determine if statistically significant differences of means exist across engineering curricula in relation to a common set of assessment criteria like ABET’s. Frequently, attitudes or behaviors are measured using five-point response scales. Familiar examples of these types of scales would include response ranges such as Strongly Agree, Agree, Neither Agree nor Disagree, Disagree, and Strongly Disagree for attitudes, or No Chance, Slight Chance, Some Chance, Good Chance, and Excellent Chance for behaviors.

When five-point, forced choice scales are used without recoding, respondents will return an answer to specific questions with means that may range from one to five numerically. If most
survey subjects respond similarly to a question, say agreeing or disagreeing, there will be very little variation on the scale. If, on the other hand, differently classified respondents (say mechanical engineering and electrical engineering students) reacting to the same survey question divide themselves into two camps, roughly half strongly agreeing and roughly half strongly disagreeing with the item, variation will approach the maximum (a bimodal distribution anchored by the extreme values on the scale).

The ability to detect statistically significant mean differences is directly related to the variance associated with responses to the survey question and the number of respondents. Everything else being equal, the capacity to detect statistically significant mean differences increases as the number of respondents does. One practical problem for assessment is that all too often the number of respondents is insufficient to allow common statistical tests to detect statistically significant differences of means with much confidence. A related analytical problem appears when the standard deviation of all responses to a survey question is so large that the number of respondents must be impractically large for the statistical test to enable us to affirm the statistical significance of observed differences that evaluators consider large enough to be practically significant.

Some may argue that one way out of the dilemma is to employ a simple chi-square test to assess the statistical significance of the deviations of the response patterns to a survey question from a “naïve” distribution (all possible responses equally likely). Another approach would be to use an empirically determined response pattern, for example one value of a nominal level of measurement contrast variable, like engineering major or class level, compared to another. However, several problems exist with either strategy. Using the naïve distribution of responses for comparative purposes involves affirming the unrealistic assumption that all values on the scale are, in truth, equally likely responses. This is rarely the case in survey-based assessment research because most students answering a survey question will typically evaluate the specific educational experience enshrined by the question favorably or unfavorably. Therefore, comparing a skewed distribution of responses to a survey item to an artificial, equal frequency distribution yields no analytically useful information despite any statistical significance of the results. Similarly, assuming that the frequency distribution for one group of respondents should be like that of another group puts the empirical cart before the horse and, even more importantly, fails to consider what the average magnitude of the observed differences are. A far better approach, one that is also much more convincing to on-campus and external constituencies, is to move away from the nominal level of measurement and to analyze assessment data using more powerful interval level statistical tests. The independent samples difference of means $t$-test is a good choice for comparing the mean responses of different contrast groups of respondents on a survey question-by-question basis.

Power analysis helps us to determine how many students we must survey when the level of confidence varies (Silverman)\(^1, 2 \& 3\). Power is a critically important but all too frequently neglected aspect of survey research design. Unless adequate statistical power is allowed for in the design of a study and the distribution of the instrument, a survey could be employed that generates too few respondents per cell (mutually exclusive groups of respondents however defined) to permit the detection of any pre-determined, substantively important differences in the population of interest. For market researchers, considering power is crucial because market
research is expensive and the results are expected to generate valid statistical inferences that facilitate the optimization of marketing strategy. In direct marketing applications, due to the large number of prospects solicited, even very small differences in the likelihood of response may translate into very large disparities in profitability. Similarly, assessment research is expensive and it should have some chance of providing meaningful feedback that identifies areas of comparative programmatic strength and weakness. In either context, the ability to say that we have high confidence that the differences observed across any contrast categories related to the population of interest are due to chance or non-chance factors is of the utmost importance to statistical conclusion validity.

The first expository task is to approximate the range of standard deviations that we might expect to see derived from responses to a survey question that uses a five-point scale. Let’s look at several extreme cases first. The first is a perfect bimodal distribution where 50 of 100 respondents Strongly Agreed (coded 5) with a survey item while 50 other respondents Strongly Disagreed (coded 1). For this distribution the mean is 3.0 and the standard deviation is 2.0. A similarly unlikely case arises if we consider a distribution where 50 respondents Agreed (coded 4) and 50 subjects Strongly Agreed (coded 5) to a survey question: for this distribution the mean is 4.5 and the standard deviation is .5. However, in the real world of higher education program assessment we are unlikely to encounter standard deviations as large or small as those noted above. Considering the results from our various student assessment surveys that use five-point response scales, the standard deviations observed typically range from about 1.0 to 1.4. Examining this range of standard deviations should prove useful to the reader who is interested in conducting assessment research using surveys that have a fighting chance of detecting statistically significant differences.

The first task for the assessment team is to decide how big a difference their instrument must be able to detect. This decision underscores the relationship between statistical power and the level of confidence. Most assessment professionals would want to be able to detect a half-point difference arising from the responses of two different groups to a survey question on a five-point scale. The conventional approach utilized in standard statistical testing sequences is to work at the 95 percent level of confidence, equivalent to one chance in 20 or less of making a Type I error (incorrectly concluding that there is a real difference when there is not). However, as Table 1 shows, by choosing a statistical significance level of .05, the assessment team may hamstring its ability to detect a large and substantively significant difference of a half point on a five-point scale if variance (the square of the standard deviation) in response to the question is high or the number of respondents is insufficient.
Table 1

The probability of being able to detect a statistically significant half-point difference of means in responses to an assessment survey question using a five-point scale, by minimum number of respondents in the cell, the population standard deviation, and the level of confidence

<table>
<thead>
<tr>
<th>Minimum in cell</th>
<th>50</th>
<th>75</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard dev.</td>
<td>1.0</td>
<td>1.2</td>
<td>1.4</td>
</tr>
<tr>
<td>1.0</td>
<td>.70</td>
<td>.86</td>
<td>.94</td>
</tr>
<tr>
<td>1.2</td>
<td>.54</td>
<td>.71</td>
<td>.83</td>
</tr>
<tr>
<td>1.4</td>
<td>.43</td>
<td>.59</td>
<td>.72</td>
</tr>
<tr>
<td>Confidence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.05</td>
<td>.80</td>
<td>.92</td>
<td>.97</td>
</tr>
<tr>
<td>.10</td>
<td>.67</td>
<td>.82</td>
<td>.90</td>
</tr>
<tr>
<td>.20</td>
<td>.69</td>
<td>.96</td>
<td>.99</td>
</tr>
<tr>
<td></td>
<td>.79</td>
<td>.90</td>
<td>.95</td>
</tr>
<tr>
<td></td>
<td>.69</td>
<td>.82</td>
<td>.89</td>
</tr>
</tbody>
</table>

Adapted from (Silverman)³.

The interpretation of Table 1 is straightforward. If you are comparing the responses of two groups of students, say men and women, on an assessment survey question the first thing to do is to determine how many respondents are in the smallest cell. Then calculate the standard deviation for all respondents (using Excel or a statistical software package unless you are unusually patient). Finally, determine what level of confidence you can live with so you don’t conclude incorrectly (the confidence level in Table 1 shows the probability of making a Type I error) that a half-point mean difference between men and women exists when it does not. Thus, if you have 50 women in the smallest cell and the standard deviation is 1.4 and you want to make an incorrect inference one time in 20 or less (the .05 level of confidence), the probability that you will be able to discern a half-point difference between men and women is only .43. This means that 57 percent of the time if a “real” half-point difference exists between men and women you will not detect it! Even if you relax the level of confidence to .10 (allowing yourself to conclude incorrectly that a half-point difference exists one time in ten), with 50 cases in the smallest cell and a standard deviation of 1.4, you could expect to detect a “real” half-point difference of means between the genders on the question only 56 percent of the time! To put it another way, the interaction of the minimum cell size, the population standard deviation, and the level of confidence equates to 56 percent power to detect a half-point difference on a five-point scale.

A second glance at Table 1 clearly shows how power improves as the minimum cell size increases, the population standard deviation decreases, and the level of confidence is relaxed. However, the important point in the assessment feedback context is that it is difficult if not impossible for many, if not most, assessment surveys to ascribe statistical significance to detected differences that most stakeholders would agree are substantively significant in practice.

Table 2 shows how rapidly minimum cell sizes must increase if detecting a quarter-point difference is the objective. To double the precision while holding the population standard deviation and statistical power constant, the minimum cell size must increase by a factor of four. The small differences seen in the far right columns of tables 1 and 2 are due to computational rounding.
Table 2

The probability of being able to detect a statistically significant quarter-point difference of means in responses to an assessment survey question using a five-point scale, by minimum number of respondents in the cell, the population standard deviation, and the level of confidence.

<table>
<thead>
<tr>
<th>Minimum in cell</th>
<th>200</th>
<th>300</th>
<th>400</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard dev.</td>
<td>1.0 1.2 1.4</td>
<td>1.0 1.2 1.4</td>
<td>1.0 1.2 1.4</td>
</tr>
<tr>
<td>Confidence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.05</td>
<td>.70 .54 .43</td>
<td>.86 .72 .59</td>
<td>.94 .84 .71</td>
</tr>
<tr>
<td>.10</td>
<td>.80 .67 .56</td>
<td>.92 .82 .71</td>
<td>.97 .90 .81</td>
</tr>
<tr>
<td>.20</td>
<td>.89 .79 .69</td>
<td>.96 .90 .82</td>
<td>.99 .95 .89</td>
</tr>
</tbody>
</table>

Adapted from (Silverman)3.

For all but the largest engineering programs, the prohibitively high numbers of survey respondents needed to characterize small differences of means as statistically significant is a meaningful analytical constraint. If the population standard deviations are lower than those used in the tables, of course, smaller differences of means will be detectable at the same level of confidence using the same minimum cell sizes. However, highly favorable conditions are not ones that should be counted on whenever student assessment survey projects are in the planning phase because the population standard deviations associated with any survey instrument are impossible to predict a priori and will vary, sometimes greatly, by question. Moreover, lower than expected response rates to an instrument will often frustrate efforts to build minimum cell sizes into the design even in the unlikely but fortuitous event that the population standard deviations for all the questions asked proved reasonably estimable. Even so, by using either of the tables above as a rough guide, assessment researches are better equipped to design their studies to have a chance of detecting the statistically significant mean differences they need to.

Implications for Completing the Feedback Loop

Most stakeholders value statistically significant results. And they should. To say that something is higher or lower than something else without a generally accepted frame of reference is unsatisfactory if not disingenuous. Statistical tests allow us to conclude, with more or less confidence, whether the differences we observe across contrast variables are likely to be real or due to random chance. These tests are able to do this because they are able to discern to what extent independent response distributions overlap. Depending on the standard deviations of the distributions and the number of cases, identical differences of means observed between two groups may be statistically significant (more likely to be real) or non-significant statistically (more likely to be due to chance). Importantly, whenever assessment feedback is employed to make policy decisions that potentially affect students or faculty members or anyone else adversely, the stamp of statistical significance reduces the likelihood that decisions will be made arbitrarily.
In the educational assessment context, the previous four pages have a nihilistic flavor. They do so because of the frequent misuse of assessment results. In their zeal to run any evidence of improvement up the flagpole, many program evaluators seize on any favorable finding. This is the wrong approach because it fails to consider the regression of short-term sample means to the long-term population mean. Indicators that go up one semester will likely fall the next. Unless statistical tests are made there is no way to estimate the likelihood that any observed changes are due to chance or actual program improvement.

The key for most colleges and departments to enhance the credibility of their assessment efforts is to bring statistical rigor into the process. As the tables demonstrate, the biggest problem for most programs using surveys as one of their primary assessment strategies is that they often collect too few responses to employ statistical tests in a meaningful way. Several data pooling and response enhancing strategies that exist for boosting the number of respondents are discussed below.

**Use Captive Audiences:** Whenever possible, survey assessment should take advantage of easily accessible respondents. When in doubt, field the instrument in class. If you desire to survey graduating seniors, survey them before they graduate and scatter to the four winds. If you assess the capstone experience, distribute the instrument at a time and place guaranteed to reach participants. If you conduct a seminar and want experiential feedback from it, distribute and collect the instrument at the event. If you want to field a survey to incoming freshmen, set up an orientation session and distribute the survey during it. While it is fine to use electronic and direct mail for follow-up, depending on these as your primary ways of reaching prospects will lead to disappointing response rates and lots of questions. If your response rate is in the single digits, it’s very hard to maintain that the results are representative of the target population because it will be impossible to gainsay the assertion that non-respondents differ from respondents in systematic yet unknown ways. And even if the estimates are statistically adjusted to mitigate the impact of a low response rate, the validity of the procedure will be lost on large segments of the audience, particularly those hostile to the findings, at multiple points in the feedback loop.

**Programmatic Assessment:** Course-level survey assessment will rarely boast numbers of respondents that are adequate for the detection of mean differences that evaluators agree have practical significance. Probably the best way to increase the number of cases available for analysis is to consider individual courses as elements within an engineering program. This approach is also consistent with ABET’s EC2000 assessment approach that uses the program as the primary unit of analysis. Since the coverage of ABET a-k criteria varies considerably from class to class, only at the academic program level is coverage of the multifarious learning criteria likely to be comprehensive. One of the greatest benefits of programmatic assessment is that engineering schools or colleges may use survey feedback to assess student learning outcomes in a comparative context. If, for example, students in one program report significantly higher criterion-related learning outcomes relative to another, and there is no reason to expect large differences across the programs due to curricular dissimilarity, a useful dialogue may begin between the departments that helps to identify what the leading department is doing that the lagging one is not. Assuming the incorporation of the leading department’s practices proves appropriate and feasible for the laggard, improvement of the program is the likely result of such feedback.
Longitudinal Assessment: One of the cornerstones of the ABET EC2000 approach is to use assessment feedback as a means for continuously improving engineering programs. To accomplish this, the usage of tests of statistical significance to demonstrate program improvement over time may be of great utility. Whether aggregate or individual level data is used to summarize program improvement, all statistical tests are sensitive to changes in their denominators. Therefore, the longitudinal pooling of assessment data by entering cohort, academic year, semester, course, course sequence, program, or other indicators of academic status and performance that make sense for the institution facilitates statistical interpretation by increasing the number of cases and concomitantly increasing the power of every statistical test to detect practically significant differences. In addition to being able to answer questions related to the magnitude and direction of change, the usage of statistical testing also permits evaluators to consider whether any observed, apparent changes in magnitude or direction over time are real or illusory. The capacity to generate longitudinal, programmatic feedback is the heart of the assessment process.

Policy Assessment: Often assessment feedback results in policy change. However, quite frequently the underlying causal relationship between a policy change and the set of conditions and behaviors that the change in policy is supposed to influence are unclear. For example, let’s say an engineering college detected what appears to be a downward trend in student course evaluations as they relate to ABET-defined learning outcomes over several semesters. Let’s also say that the student-instructor ratio increased during the period. Noting the apparent correlation and having little else to go on, but without conducting any statistical analysis, the college administration caps course enrollments at levels that existed prior to the observed decline in performance. Feedback from the next few assessment cycles is inconclusive. Student performance seemed to increase the first semester after the student-faculty ratio was reduced but then performance declined the next two terms. While this is a simplistic example of what may happen when decision-makers take full advantage of chance, the important point is that by pooling course level enrollment and performance data and conducting simple statistical tests the college could easily have determined whether the observed declines were likely due to chance or unlikely enough to necessitate a change in policy. By testing for rather than assuming causation, the assessment feedback looping back into the decision-making process would have been filtered though the lens of statistical confidence before consumption. The result would likely have been, in this instance, an appropriate non-decision.

Database Design and Development: How an institution organizes its assessment data determines what it is able to do with it. A relational database system holding all manner of individual level, historical, student and institutional data accessible through query language is clearly the optimal design. Using individual level keys, files may be merged, variables extracted into temporary data sets, and all manner of useful descriptive, explanatory, and predictive analyses may be conducted. Unfortunately, at some institutions the debates over adopting comprehensive relational database systems were won by fiscal conservatives or atavists desiring to maintain their influence by limiting widespread access to informational resources. Incompatible data storage and retrieval systems are often the result. Sometimes the need to enhance response rates to assessment instruments by preserving the anonymity of respondents,
thus preventing the merging of data records from multiple sources, is controlling. In either case, the potential utility of assessment feedback is constrained.

Under constraint, in a flat file environment, the assessment team must carefully consider its options. The easiest way to organize the survey-derived data is by instrument. Since surveys should evolve over time (each instrument fielded is a pilot for the next iteration), a variable field should be retained for each question fielded in perpetuity. Categorical contrast variables summarizing the year, semester, academic department, instructor, and course, at a minimum, should be incorporated into each data matrix. Rewording and rescaling of good questions should be kept to a minimum to reduce the potential for confounding instrumentation effects. Students should be asked to provide information related to their satisfaction with their expected grades in the course, class levels, and the like to facilitate differential assessment of the responses of differently situated groups. Across the surveys, common questions should appear so that aggregates may be easily compared longitudinally, programmatically, and from instrument to instrument. If multiple flat files are used it’s more difficult to make sense of the totality of assessment results because the capacity for analysis is limited and the process takes more steps. Even so, no matter what database structures are employed, without an adequate number of cases to detect practically significant differences the assessment process will resemble a highly dubious exercise: feeling around in a dark room for the light switch.

Conclusions

This brief paper shows that no matter how well assessment feedback is incorporated into the process, the potential for assessment data to provide useful information to better inform educational policy decisions may be hindered by very basic research design issues. The inability to detect practically significant real differences across groups of respondents to assessment surveys may lead to inaccurate programmatic inferences and ill-informed decisions. Moreover, without the stamp of statistical significance evaluators may have a difficult time demonstrating improvement to external consumers of assessment data such as senior administrators, university councils, state boards of higher education, program review committees, and the like. By using captive audiences and employing programmatic and longitudinal analysis in their assessment plans, academic units will be better positioned to have the number of respondents they need to make a compelling statistical case for their survey results.

What this paper does not argue for is the wholesale abandonment of non-statistical modes of assessment inquiry. Clearly, the best approach is to obtain programmatic data from multiple sources. Qualitative methodologies such as focus groups, in-depth interviewing, expert appraisals, and feedback from both program faculty and external clients of various types are extremely valuable data resources that should be incorporated in whole or in part into every comprehensive assessment plan. Equally valuable are assessment data derived from the entering characteristics of students, performance on standardized tests, portfolio assessment, and demonstration projects. Similarly, aggregate institutional data related to program rankings in national publications, student enrollment, time to degree, and retention and graduation rates have their place in every comprehensive assessment plan. By triangulating survey, qualitative, and institutional performance data, academic programs will be best positioned to make sense of the subtleties, complexities, and contradictions inherent in the assessment of ongoing student
learning. By incorporating multiple feedback channels into the assessment process, an academic program will more easily achieve a condition of dynamic, self-regulating, homeostatic equilibrium with its environment. Designing survey research so that it has a fighting chance of fulfilling its potential is an important consideration for any comprehensive assessment plan.

**Bibliography**


**ED FURLONG**
Furlong is the Assessment Coordinator for the College of Engineering and Engineering Technology at Northern Illinois University. Furlong received his Ph.D (Political Science-Public Policy Analysis) from NIU.

**PROMOD VOHRA**
Vohra is the Associate Dean of Northern Illinois University’s College of Engineering and Engineering Technology. Vohra received both his Ed.D (Instructional Technology) and M.S. (Electrical Engineering) from NIU.